LREC 2026

Joint Workshop on Readability and Text Simplification (READIxTSAR) @ LREC 2026

Workshop Proceedings

Editors
Matthew Shardlow, Thomas François, Raquel Amaro, Jorge Baptista, Rémi Cardon, Eugénio Ribeiro, Horacio Saggion, Regina Stodden, Amalia Todirascu, and Rodrigo Wilkens

11 May 2026

Proceedings of the Joint Workshop on Readability and Text Simplification (READIxTSAR) @ LREC 2026

ISBN 978-2-493814-91-3

Preface

Recent studies show that the number of children and adults facing difficulties in reading and understanding written texts is steadily growing. Reading challenges can show up early on and may include reading accuracy, speed, or comprehension to the extent that the impairment interferes with academic achievement or activities of daily life. Various technologies (text customization, text simplification, text-to-speech devices, and screening for readers through games and web applications, to name a few) have been developed to help poor readers to get better access to information as well as to support reading development. Among those technologies, text adaptations are a powerful way to leverage document accessibility by using NLP techniques.

Thus, the growth of educational and assistive technologies for reading, aimed at enhancing the performance of individuals with reading difficulties, intellectual disabilities, or reading in a second language, provides an important setting for automated text simplification (ATS) research, which is one the key strategies for enhancing text accessibility. Nevertheless, ATS research often remains insufficiently grounded in the practical realities of education and disability contexts, thereby reducing the number of evaluations conducted with these target groups and constraining the integration of theoretical principles from education and accessibility research into ATS systems. This READIxTSAR workshop aims to present state-of-the-art applications and approaches in technology-enhanced reading along with innovations in text accessibility with the aim of bringing together these two strands of research. The workshop addresses specialized technology, tools, and resources, their impact on learning to read and comprehension, and innovative works spanning research to fieldwork, particularly in light of recent AI advances.

READIxTSAR is co-located with LREC 2026, the 15th edition of the Language Resources and Evaluation Conference, held in Palau de Congressos de Palma, Palma de Mallorca (Spain), on 11-16 May 2026. READIxTSAR is a joint initiative between two previous workshops of mutual interest: Tools and Resources for REAding DIfficulties (READI, hosted at LREC 2020,22,24) and Text Simplification, Accessibility and Readability (TSAR, EMNLP 2022,24,25, RANLP 2023). This year at LREC, the organising committee and programme committee of the two events have merged to deliver a joint event, uniting accessibility research communities under a common umbrella. Whereas the TSAR community is more oriented towards text simplification technology and the production of NLP methods for advancing the state of the art in text simplification, the READI community is oriented towards the development of novel tools and resources for improving the understandability of written documents. By uniting these two communities for a collaborative event, we hope to share mutual experiences and best working practices in our fields of interest fostering new collaborations and developing capacity for future research avenues of mutual benefit.

We received 24 submissions to the workshop, of which none were desk rejected. All 24 submissions were submitted for peer review, with each submission receiving two or three peer reviews from the programme committee, with papers aligned to reviewer expertise. Additionally, each paper was assigned to an expert member of the organising committee to coordinate reviews and make a provisional decision. All decisions were ratified by the full organising committee, with a total of 17 papers accepted of which 6 papers were accepted as oral presentation and 11 as poster presentations. This gives an acceptance rate of 70.83%, which is in line with previous editions of READI and TSAR. The organising committee aimed to offer developmental feedback based on reviewer suggestions to ensure that camera ready versions of all papers were of high quality when presented to the workshop audience.

The authorship of the accepted works are geographically diverse, with authors from across

Europe and beyond (Switzerland, France, Belgium, Portugal, UK, Spain, The Netherlands, Italy, Germany, Japan, UAE, USA). Topics of the accepted papers ranged from: novel methods and evaluation metrics for aspects of text simplification (6 papers); novel studies on the nature of readability including LLM-mediated texts (6 papers); studies at the intersection of simplification and readability investigating the interplay between these related phenomena (2 papers); and studies on representational aspects of complexity in models and embeddings (3 papers). The papers also exhibited linguistic diversity in the texts studied and evaluated. Whereas English was the primary language of study for 9 accepted papers, submissions also were accepted which studied French (3), Portuguese (2), Arabic (1), German (1) and Galician/Spanish (1).

The organisers extend their thanks to the authors who kindly submitted their work to our workshop, and to our program committee members, the reviewers and the additional reviewers who did a thorough job evaluating submissions. We would also like to acknowledge the participants of the workshop who attended both the oral and poster sessions. We would like to thank the committee of the DeTermIt workshop, with whom we co-ordinated our programme to ensure that participants could benefit from a full day of presentation of simplification and readability related presentations.

The organisers also wish to acknowledge the participation of the iRead4Skills and iDEM projects within the workshop, both of which provided an invited speaker to discuss the accessibility focussed research arising from each project.

Finally, we wish to acknowledge the work of the LREC organising committee and particularly the workshop chairs in selecting and supporting our workshop in the LREC 2026 programme of events.

The READIxTSAR Organising Committee

Organising Committee

- Matthew Shardlow, Manchester Metropolitan University, UK
- Thomas François, UCLouvain, Belgium
- Raquel Amaro, NOVA University Lisbon, Portugal
- Jorge Baptista, Universidade do Algarve & INESC-ID Lisboa, Portugal
- Rémi Cardon, Universidad Carlos III de Madrid, Spain
- Eugénio Ribeiro, Iscte-IUL & INESC-ID Lisboa, Portugal
- Horacio Saggion, Universitat Pompeu Fabra, Spain
- Regina Stodden, University Bielefeld, Germany
- Amalia Todirascu, Université de Strasbourg, France
- Rodrigo Wilkens, University of Exeter, UK

Table of Contents

Revisiting German Complex Word Identification: Contextualized LLMs and Feature Injection
Thorben Schomacker, Seid Muhie Yimam, Chris Biemann and Marina Tropmann-Frick . . 1

Book Complexity Level Assignment in French and Portuguese
Jorge Baptista, David Antunes, Wafa Aissa, Julien Zakhia Doueihi, Hanh Trang Tran Pham, Eugénio Ribeiro, Thomas François and Raquel Amaro 12

Taming CATS: Controllable Automatic Text Simplification through Instruction Fine-Tuning with Control Tokens
Hanna Hubarava and Yingqiang Gao 26

PLABA-EVAL: A Multi-Dimensional, In-Context Sentence Readability Dataset for Medical Text
Kexin Bian, Su-Youn Yoon and Mamoru Komachi 49

Automatic Extraction of Textual and Phonemic Complexity for French Cued Speech
Magali Norré, Brigitte Bigi, Núria Gala, Ludivine Javourey Drevet and Thomas François 61

Can LLMs Control Readability? A Multi-Dimensional Evaluation Framework for CEFR-Controlled Arabic Generation
Nour Rabih, Chatrine Qwaider and Ted Briscoe 74

Lexical Conditioning of Model's Distribution through Uncertainty-gated Soft-Mixing of Probabilities
Michele Papucci, Giulia Venturi and Felice Dell'Orletta 89

A Comparative Study of Multilingual Fine-tuning and Prompting for Automatic Text Readability Classification in Galician
Sandra Rodríguez Rey and Marcos Garcia 101

Plan-Guided Text Simplification with Extended Contexts
Pascal Mathas, Jan Bakker and Jaap Kamps 121

LLM-Generated Stories for Students with Significant Cognitive Disabilities: Promise, Gaps, and Evaluation Framework
Pragati Maheshwary, Ananya Ganesh and Shamya Karumbaiah 130

Evaluating Transformer Model Family Representations Through Automated Essay Scoring
Akchay Ozten and Rodrigo Wilkens 142

Proficiency-Controlled Text Simplification in European Portuguese: A Preliminary Study using Prompting Approaches
Eugénio Ribeiro, David Antunes, Nuno Mamede and Jorge Baptista 151

Automatic Text Simplification for French Medical Documents with LLMs: The Role of Target Audience and Genre
Rémi Cardon and A. Seza Dogruoz 164

A Learner-Oriented Annotated Resource of French Multiword Expressions for Text Adaptation in Foreign Language Reading
Anna Kalinina, Thomas François, Hélène Vassiliadou and Amalia Todirascu 181

A Meta-evaluation of Automatic Metrics for Elaborative Simplification
Abdullah Alshatti, Steven Schockaert and Fernando Alva-Manchego 193

Readability Measures in Automatic Text Simplification: Is Simplification Quality a Coherent Construct?
Rémi Cardon and A. Seza Dogruoz 210

Language Proficiency as a Recoverable Dimension in Multilingual LLM Embeddings
Rodrigo Wilkens 227

Workshop Program

Monday, May 11, 2026

09:00–09:10 **Welcome Session**

09:10–10:00 **Keynote Session**

09:10–09:35 *The iRead4Skills Project*
Raquel Amaro

09:35–10:00 *The iDEM Project*
Horacio Saggion

10:00–10:30 **Oral Presentations I**

10:00–10:15 *Revisiting German Complex Word Identification: Contextualized LLMs and Feature Injection*
Thorben Schomacker, Seid Muhie Yimam, Chris Biemann and Marina Tropmann-Frick

10:15–10:30 *Book Complexity Level Assignment in French and Portuguese*
Jorge Baptista, David Antunes, Wafa Aissa, Julien Zakhia Doueihi, Hanh Trang Tran Pham, Eugénio Ribeiro, Thomas François and Raquel Amaro

10:30–11:00 **Coffee Break**

Monday, May 11, 2026 (continued)

10:30–11:30 **Poster Session**

10:30–11:30 *Taming CATS: Controllable Automatic Text Simplification through Instruction Fine-Tuning with Control Tokens*
Hanna Hubarava and Yingqiang Gao

10:30–11:30 *PLABA-EVAL: A Multi-Dimensional, In-Context Sentence Readability Dataset for Medical Text*
Kexin Bian, Su-Youn Yoon and Mamoru Komachi

10:30–11:30 *Automatic Extraction of Textual and Phonemic Complexity for French Cued Speech*
Magali Norré, Brigitte Bigi, Núria Gala, Ludivine Javourey Drevet and Thomas François

10:30–11:30 *Can LLMs Control Readability? A Multi-Dimensional Evaluation Framework for CEFR-Controlled Arabic Generation*
Nour Rabih, Chatrine Qwaider and Ted Briscoe

10:30–11:30 *Lexical Conditioning of Model's Distribution through Uncertainty-gated Soft-Mixing of Probabilities*
Michele Papucci, Giulia Venturi and Felice Dell'Orletta

10:30–11:30 *A Comparative Study of Multilingual Fine-tuning and Prompting for Automatic Text Readability Classification in Galician*
Sandra Rodríguez Rey and Marcos Garcia

10:30–11:30 *Plan-Guided Text Simplification with Extended Contexts*
Pascal Mathas, Jan Bakker and Jaap Kamps

10:30–11:30 *LLM-Generated Stories for Students with Significant Cognitive Disabilities: Promise, Gaps, and Evaluation Framework*
Pragati Maheshwary, Ananya Ganesh and Shamya Karumbaiah

10:30–11:30 *Evaluating Transformer Model Family Representations Through Automated Essay Scoring*
Akchay Ozten and Rodrigo Wilkens

10:30–11:30 *Proficiency-Controlled Text Simplification in European Portuguese: A Preliminary Study using Prompting Approaches*
Eugénio Ribeiro, David Antunes, Nuno Mamede and Jorge Baptista

10:30–11:30 *Automatic Text Simplification for French Medical Documents with LLMs: The Role of Target Audience and Genre*
Rémi Cardon and A. Seza Dogruoz

Monday, May 11, 2026 (continued)

11:30–13:00 **Oral Presentations II**

11:30–11:50 *A Learner-Oriented Annotated Resource of French Multiword Expressions for Text Adaptation in Foreign Language Reading*
Anna Kalinina, Thomas François, Hélène Vassiliadou and Amalia Todirascu

11:50–12:10 *A Meta-evaluation of Automatic Metrics for Elaborative Simplification*
Abdullah Alshatti, Steven Schockaert and Fernando Alva-Manchego

12:10–12:30 *Readability Measures in Automatic Text Simplification: Is Simplification Quality a Coherent Construct?*
Rémi Cardon and A. Seza Dogruoz

12:30–12:50 *Language Proficiency as a Recoverable Dimension in Multilingual LLM Embeddings*
Rodrigo Wilkens

12:50–13:00 **Closing Session**

Revisiting German Complex Word Identification: Contextualized LLMs and Feature Injection

Thorben Schomacker[1,2], **Seid Muhie Yimam**[1], **Chris Biemann**[1], **Marina Tropmann-Frick**[2]
[1]University of Hamburg, [2]Hamburg University of Applied Sciences
thorben.schomacker@haw-hamburg.de

Abstract

Complex word identification (CWI) is essential in text simplification, yet work on German CWI remains comparatively limited. To address this gap, we investigate the capabilities of three state-of-the-art LLMs and compare them to previously proposed baseline systems. We fine-tune the LLMs in three setups: (i) using the target expression only, (ii) using the target expression together with its sentence-level context, and (iii) using the context and injection of classical machine learning features. Our results show that while pretrained-only LLMs fall short, fine-tuned LLMs set new benchmarks for both binary and probabilistic CWI. In addition, embedding the target in its context sentence improves performance, whereas feature injection has no clearly measurable effect. All models in this paper are trained on the probabilistic CWI task and additionally evaluated on the binary task; thus, we publish a single model that supports both evaluation views.

We released all accompanying resources (https://github.com/tschomacker/german-cwi-llm) *and model checkpoints* (https://huggingface.co/collections/tschomacker/german-cwi-llm).

Keywords: Complex Word Identification, Lexical Simplification, Sequence Classification, German Data

1. Introduction

This paper evaluates the capabilities of state-of-the-art LLMs (two LLaMA-style models and one BERT-based model) on German complex word identification (CWI). CWI is a key prerequisite for lexical and text simplification, as it identifies the units that are most likely to block comprehension for a given audience and therefore should be prioritized for simplification or explanation. We compare our results to 14 baseline systems previously reported for the same dataset and task setting.

In addition to the neural models, we consider length- and frequency-based features, as they have been widely used in CWI systems and are reported to be robust cross-linguistic predictors of lexical complexity (Bingel and Bjerva, 2018). More generally, the question of which features best capture word complexity has been investigated in a range of studies (see Gooding (2023, Section 2.1.3) for an overview).

Many CWI approaches operate in a context-independent setting by predicting complexity from the target expression alone, without explicitly modeling the sentence in which it occurs. However, CWI can also be formulated in ways that incorporate context information, for example as a sequence labeling problem (Gooding and Kochmar, 2019). In this work, we explicitly test whether sentence-level context improves German CWI performance when training and evaluating LLMs.

Recent LLM-based approaches have reported strong results on English CWI but more limited performance for German. Although prior work suggests that LLMs may show only modest performance for CWI (Smădu et al., 2024), we evaluate more recent models and training setups to assess current capabilities. We address the following research questions:

RQ1 Does sentence-level context information affect the performance of LLMs on German CWI?

RQ2 Do explicitly injected statistical features provide additional benefit beyond contextualized transformer representations?

RQ3 How sensitive is binary complex word identification performance to the choice of decision threshold when using probabilistic modeling?

2. Related Work

Complex word identification (CWI) is a central component of lexical and automatic text simplification pipelines (Shardlow, 2015, p. 29). While extensively studied for English, German CWI has received comparatively limited attention. Early approaches largely mirrored English work (e.g., Shardlow, 2013) and relied on surface-level indicators such as word length, syllable count, and corpus frequency.

A major step forward was the CWI 2018 Shared Task (Yimam et al., 2018), which released the first publicly available German portion of a multilingual CWI dataset annotated by L2 speakers (Yimam et al., 2017). The dataset provides both binary and probabilistic word-level complexity judgments and

Proceedings of the Joint Workshop on Readability and Text Simplification (READIxTSAR) @ LREC 2026, pages 1–11
11 May 2026.

Sentence	*native* **total**	*non-native* **total**	*native* **complex**	*non-native* **complex**	*complexity* **Bin.**	**Prob.**
An der Maschine wurde das <target dwds=2 len=8 syl=2 vow=2>Fahrwerk</target> beschädigt. *The machine's <target dwds=2 len=8 syl=2 vow=2>landing gear</target>* was damaged.	3	9	0	2	1	0.167
An der <target dwds=4 len=8 syl=3 vow=3>Maschine</target> wurde das Fahrwerk beschädigt. *The <target dwds=4 len=8 syl=3 vow=3>machine's</target> landing gear was damaged.*	3	9	0	0	0	0.0
Hauptgrund für die Verschlechterung des Zustandes sei der heiße und trockene Sommer 2003 mit hohen <target dwds=1 len=10 syl=3 vow=4>Ozonwerten</target> . *The main reason for the deterioration in the situation is said to be the hot and dry summer of 2003 with high <target dwds=1 len=10 syl=3 vow=4>ozone levels</target> .*	6	6	5	5	1	0.833
Hauptgrund für die <target dwds=None len=30 syl=8 vow=8>Verschlechterung des Zustandes</target> sei der heiße und trockene Sommer 2003 mit hohen Ozonwerten. *The main reason for the <target dwds=None len=30 syl=8 vow=8>deterioration in the situation</target> is said to be the hot and dry summer of 2003 with high ozone levels .*	6	6	0	1	1	0.083

Table 1: Examples from training data in the CWI 2018 dataset, represented using the preprocessing schema (see Section 3.1.2) we have developed. The first column is the preprocessed input sentence, third and fourth column is the total number of native/non-native annotators and fifth and sixth column refer to the number of annotators, who annotated the target expression as complex. **Bin.** is the binary complexity label, which is one if at least one annotator rated the target as complex and 0 in all other cases. **Prob.** is the probabilistic complexity score, which is the percentage of the annotators marked the word as complex. In our cases $2/12 = 0.167$, $0/9 = 0.0$, $10/12 = 0.833$ and $1/12 = 0.083$

established standard evaluation settings. Baseline systems relied primarily on frequency- and length-based features.

Among the shared task submissions, tree-based ensemble methods proved competitive. For example, Kajiwara and Komachi (2018) employed random forest classifiers and regressors with features including the number of characters, number of words and the frequency in a corpus written by native speakers and a corpus written by language learner. In particular, they use Lang-8 (Mizumoto et al., 2011), a learner corpus for eight languages (including German). Similarly, Bingel and Bjerva (2018) introduced CoastalCPH, combining an ensemble of feed-forward neural networks and random forests for the binary task, and random forest regressors for probabilistic prediction. We report these baselines and additional systems in Table 3. Taken together, this line of work establishes strong feature-based baselines for German CWI and highlights the usefulness of surface and frequency indicators.

Subsequent work strengthened feature-based baselines. Finnimore et al. (2019) compiled 25 features spanning target-level, subword-level, and sentence-level properties and conducted systematic ablations to derive compact cross-lingual feature sets. Their results show that carefully engineered features can rival or outperform many shared task submissions.

More recently, the BEA 2024 Shared Task on the Multilingual Lexical Simplification Pipeline (MLSP) reframed complexity assessment as part of a two-stage pipeline, combining lexical complexity prediction (LCP) with lexical simplification (LS) on shared targets and contexts across ten languages, including German (Shardlow et al., 2024). In contrast to earlier CWI setups, LCP targets a continuous complexity score in $[0, 1]$, aligning complexity prediction more directly with downstream simplification. While MLSP provides German instances, the labeled data is not openly released (access is restricted), which limits direct comparability with the CWI 2018 evaluation setting.

Beyond feature-based classification, Gooding and Kochmar (2019) reformulated CWI as a sequence labeling problem. Using a BiLSTM architecture with word- and character-level representations and an auxiliary language modeling objective, they demonstrate that a unified neural model can outperform task-specific systems on English data. This motivates the question of whether modern

	total	**complex**	**simple**	**mean**
Train	6151 ($\sim 78\%$)	3589 ($\sim 58\%$)	2562	0.0783
Dev	795 ($\sim 10\%$)	334 ($\sim 42\%$)	461	0.0798
Test	959 ($\sim 12\%$)	376 ($\sim 39\%$)	583	0.0746
total	7905			

Table 2: The number and ratio of instances in the German portion and the mean value for the probabilistic label in the German CWIG3G2 dataset. Dataset split is the same as for the shared task.

LLMs, as general-purpose models, can improve performance in German CWI as well.

In parallel to these developments in feature engineering and task formulation, recent work has examined whether pretrained neural models can reduce reliance on manual features and improve generalization across domains. However, initial findings suggest that pretrained-only performance remains limited in this setting (Smădu et al., 2024). For German in particular, the CWI 2018 dataset remains the only publicly available resource at the word level, although related corpora addressing subjective sentence-level complexity exist (Naderi et al., 2019; Seiffe et al., 2022) and the MLSP could be used on demand. This resource landscape motivates further investigation into how modern pretrained models perform on German CWI.

3. Methodology

We used the only openly available German CWI corpus (Section 3.1) on three German LLMs (Section 3.2) with different experimental configurations. The results are discussed in Section 4.

3.1. Data

We use the German portion of the CWIG3G2 dataset introduced by Yimam et al. (2017) and used in the CWI 2018 shared task (Yimam et al., 2018).[1] Each instance (total $= 7905$) provides a target expression (which may be a single word or a multiword expression) together with sentence context and both binary and probabilistic complexity annotations.

3.1.1. Labels

In the dataset used for this work, binary complex word labels are defined existentially: a target is annotated as complex if at least one annotator marked it as difficult, and as simple otherwise. The probabilistic labels, in contrast, encode the proportion of annotators who judged a target as complex. As these two labels capture related but distinct notions of lexical complexity, we optimize the model with respect to the probabilistic labels and treat the binary labels as a separate evaluation view rather than as a thresholded variant of the probabilistic score.

For binary evaluation, we follow the shared-task definition and use the dataset-provided existential labels, where a target is considered complex if at least one annotator marked it as difficult. We obtain binary predictions by thresholding the model's predicted score $p = \sigma(z)$ at a decision threshold τ, selected exclusively on the development set (Section 3.4). In the German CWI dataset, there is a total of 23 annotators (12 native and 11 non-native speakers) (Yimam et al., 2017). In the German portion, each target is annotated by 12 annotators, so existential complexity ("at least one annotator") corresponds to $p \geq 1/12 \approx 0.083$. Accordingly, the probabilistic labels are quantized in steps of $1/12$, and the smallest non-zero value is 0.083.

We use this label definition and thresholding procedure consistently throughout the paper.

3.1.2. Pre-processing scheme

For our model inputs, we pre-process the dataset by marking the target span in the sentence using `<target>` tags. Depending on the experimental setup, we either provide (i) the target expression without sentence context, (ii) the tagged sentence containing the target, or (iii) the tagged sentence with injected features following. Table 1 shows examples of the resulting input format. Dataset statistics and label distributions are reported in Table 2. The examples in Table 1 also illustrate the quantization of probabilistic labels (e.g., $0.083 = 1/12$ and $0.833 = 10/12$).

3.1.3. Feature Injection

Yimam et al. (2018) use three length features for their models: the number of vowels, syllables and characters and three frequency features: the frequency in Simple Wikipedia, the frequency in the paragraph, and the frequency in the Google Web 1T 5-Grams.

Similarly, we count vowels (vow), syllables (syl) using the spaCy's syllables module[2] and characters (len). For frequency, we used the "Digitales Wörterbuch der deutschen Sprache" ($dwds$, English: digital dictionary of the German language)[3]. The DWDS assigns an integer value to a word, the

[1] www.inf.uni-hamburg.de/en/inst/ab/lt/resources/data/complex-word-identification-dataset CC-BY 4.0 License

[2] https://spacy.io/universe/project/spacy_syllables

[3] https://www.dwds.de/d/api

higher the frequency the higher the value. Multiword expressions yield `None`, since they are not supported by the DWDS API (see the 4th example in Table 1).

Concretely, we use the following input templates: (i) **target-only**: `<target> T </target>`; (ii) **context**: the full sentence with `<target>` tags around the marked span; (iii) **context+features**: the tagged sentence, followed by a feature string inside the opening tag, e.g., `<target dwds=2 len=8 syl=2 vow=2>`.

3.2. Models

To select strong German-pretrained models for our CWI experiments, we rely on external German NLU benchmarks as a proxy for general classification capability. Popular benchmark suites such as GLUE (Wang et al., 2018), SUPERGLUE (Wang et al., 2019) and HELM (Liang et al., 2023) do not provide German evaluation data. We therefore use SuperGLEBer (Pfister and Hotho, 2024), a German NLU benchmark suite with 29 tasks across classification, sequence tagging, sentence similarity, and question answering.[4]

SuperGLEBer does not include CWI; we use its *classification* leaderboard solely to guide model selection. We select the top three models by mean classification performance and fine-tune them on CWIG3G2:

1. `LSX-UniWue/LLaMmlein2Vec_7B`[5] (Wunderle et al., 2025)
2. `LSX-UniWue/ModernGBERT_1B`[6] (Wunderle et al., 2025)
3. `LSX-UniWue/LLaMmlein_7B`[7] (Pfister et al., 2025)

All experiments were conducted on a Tesla V100-PCIE-16GB GPU. In the following paragraphs we discuss the (hyper-) parameters used in our experiments. We adopted Pfister and Hotho (2024)'s experimental configuration as closely as possible for our model setups to foster comparability.

Epochs Similar to Pfister and Hotho (2024) we trained our models for 5 epochs. Additionally, we measure the performance on the dev-subset (not on `test`-subset to avoid bleeding) before training to evaluate the models' performance without fine-tuning and to eventually investigate the effects of the compared training paradigms.

Learning rate Similar to Pfister and Hotho (2024) we trained our models with a learning rate of $5\mathrm{e}{-5}$.

Batch size We initially used the same batch size (8) as Pfister and Hotho (2024) but observed that larger per-device batch sizes led to numerical instability (non-finite parameters) during early training steps. We therefore used smaller micro-batches ($= 1$) with gradient accumulation ($= 8$) to improve numerical stability while keeping the effective batch size the same as Pfister and Hotho (2024).

PEFT We apply parameter-efficient fine-tuning (PEFT) (Mangrulkar et al., 2022) using Low-Rank Adaptation (LoRA) (Hu et al., 2022) and quantization (QLoRA). The LoRA configuration is defined by four parameters: the rank r, the scaling factor `lora_alpha`, the task type, and the target modules. We set $r = 8$ and `lora_alpha` $= 32$, following Pfister and Hotho (2024). The task type is sequence classification (`SEQ_CLS`), as we model CWI as a sequence classification task.

For `ModernGBERT_1B`, we apply LoRA adapters to the target modules `Wqkv`, `Wi`, and `Wo`, following the model card.[8] To ensure comparability across architectures, we map these target modules to the structurally corresponding layers in the LLaMA-style models (`LSX-UniWue/LLaMmlein2Vec_7B` and `LSX-UniWue/LLaMmlein_7B`). Concretely, the fused query–key–value projection (`Wqkv`) and the attention output projection (`Wo`) correspond to the attention projections `q_proj`, `k_proj`, `v_proj`, and `o_proj` in LLaMA-style architectures, yielding a comparable attention-focused LoRA setup across all models. For both LLaMA-style models, we perform 4-bit quantization using BitsandBytes.

3.3. Problem Formulation

We model complex word identification as a probabilistic binary classification task and train the model using binary cross-entropy loss. Given an input instance x (sentence and target expression), the model produces a single logit $z \in \mathbb{R}$, which is transformed into a probability via the sigmoid function. The model parameters θ are optimized by minimizing the binary cross-entropy between the predicted probability $p(\text{complex} \mid x)$ and the gold label y:

$$p(\text{complex} \mid x) = \sigma(f_\theta(x)) = \sigma(z) = \frac{1}{1 + e^{-z}}$$

The resulting value $p \in [0, 1]$ is interpreted as a continuous complexity score, corresponding to the es-

[4] https://lsx-uniwue.github.io/SuperGLEBer-site/leaderboard_v1, last access 08.12.25

[5] https://huggingface.co/LSX-UniWue/LLaMmlein2Vec_7B, last access 19.02.26

[6] https://huggingface.co/LSX-UniWue/ModernGBERT_1B, last access 19.02.26

[7] https://huggingface.co/LSX-UniWue/LSX-UniWue/LLaMmlein_7B, last access 19.02.26

[8] https://huggingface.co/LSX-UniWue/ModernGBERT_1B , last access 03.02.2026

timated proportion of annotators who would judge the target expression as complex.

We deliberately adopt this formulation instead of discretizing lexical complexity into multiple ordinal bins and recovering a continuous score via expected values over softmax outputs, as proposed by Smădu et al. (2024). While ordinal binning provides a robust training signal for graded lexical complexity, it introduces an additional discretization step that is not required in our setting, where the target variable is already defined as a probability. Using a sigmoid-based formulation allows us to preserve the original label semantics, avoid arbitrary bin boundaries, and maintain architectural and procedural consistency with standard complex word identification setups.

3.4. Threshold Sensitivity Analysis

Since our model predicts a continuous complexity score $p = \sigma(z)$, a binary decision rule requires selecting a threshold τ. While the annotation scheme defines complexity in existential terms (i.e., at least one annotator marked the word as difficult), the exact numerical threshold depends on the normalization of the probabilistic labels.

To assess the robustness of binary performance with respect to this decision rule, we conduct a threshold sensitivity analysis. On the development set, we evaluate a small set of fixed thresholds $\tau \in \{0.042, 0.083, 0.167\}$, corresponding to values below, equal to, and above the assumed annotator-normalized threshold. We report $F1_{macro}$ for each setting while keeping the probabilistic training procedure unchanged.

Based solely on development performance, we select a single global threshold and keep it fixed for all test evaluations. This ensures that no decision rule is optimized on the test set. Test results are reported under the selected threshold and compared to existing benchmarks.

Figure 1 reports development set performance for different threshold values and training states. The selected threshold is determined exclusively based on development $F1_{macro}$.

Selected threshold We evaluate binary CWI by thresholding the predicted probabilistic score $p = \sigma(z)$ at a fixed decision threshold τ. On the development set, we compare $\tau \in \{0.042, 0.083, 0.167\}$ and select a single global threshold by maximizing $F1_{macro}$. For the fine-tuned models (5 epochs), the best development performance is obtained with $\tau = 1/24 \approx 0.042$, and we therefore fix $\tau = 0.042$ for all subsequent test evaluations.

Note that $\tau = 0.042$ lies below the smallest non-zero gold probability step ($1/12 \approx 0.083$). We nevertheless select τ strictly by development-set macro-F_1 and keep it fixed for all test evaluations.

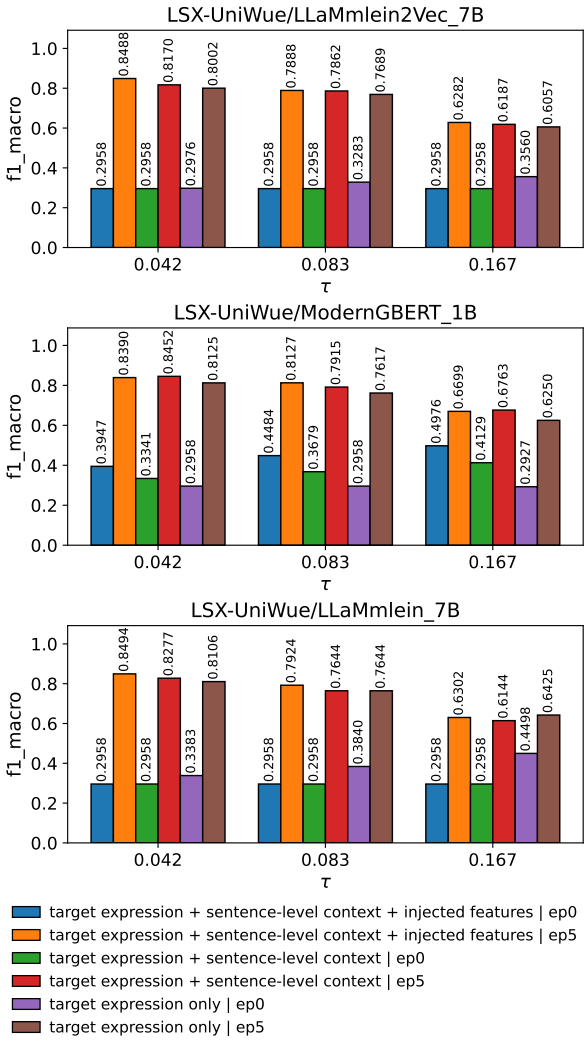


Figure 1: $F1_{macro} \uparrow$ measured on the dev-set with different setups and thresholds τ for three models. Threshold τ is selected on the development set by maximizing $F1_{macro} \uparrow$.

3.5. Evaluation

To ensure comparability, we adopt the same evaluation methodology as Yimam et al. (2018): Mean Absolute Error (MAE) for probabilistic CWI and macro-F_1 ($F1_{macro}$) for binary CWI. We train all models on the probabilistic scores and report MAE against these gold values. For binary evaluation, we follow the shared-task definition using the dataset-provided existential labels and derive binary predictions by thresholding the predicted score $p = \sigma(z)$ at the development-selected threshold $\tau = 0.042$ (Sections 3.1 and 3.4).

4. Results

Table 3 reports performance on probabilistic CWI (MAE) and binary CWI ($F1_{macro}$), where binary predictions are obtained by thresholding model

Model (sorted by publication date)					$F1_{macro}\uparrow$ DEV	$F1_{macro}\uparrow$ TEST	$MAE\downarrow$ DEV	$MAE\downarrow$ TEST
Kajiwara and Komachi (2018): TMU (CWI2018 1st rank)					-	0.7451	-	<u>0.0610</u>
Bingel and Bjerva (2018): CoastalCPH					-	0.6619	-	0.0747
Yimam et al. (2018): baseline					-	<u>0.7546</u>	-	0.0816
Finnimore et al. (2019): monolingual baseline-2-features					**0.795**	0.724	-	-
Finnimore et al. (2019): monolingual full-25-feature set					0.746	0.748	-	-
Finnimore et al. (2019): cross-lingual selected-5-features (EN+ES)					0.783	0.734	-	-
Finnimore et al. (2019): cross-lingual feature selection (ES)					0.774	0.726	-	-
Finnimore et al. (2019): cross-lingual feature selection (EN)					0.760	0.730	-	-
Smădu et al. (2024): Llama-2-7b-ft					-	0.705	-	-
Smădu et al. (2024): Llama-2-13b-ft					-	0.708	-	-
Smădu et al. (2024): Vicuna-v1.5-7b-ft					-	0.675	-	-
Smădu et al. (2024): Vicuna-v1.5-13b-ft					-	0.700	-	-
Smădu et al. (2024): Llama-3-8b-ft					-	0.708	-	-
Smădu et al. (2024): ChatGPT-3.5-turbo-ft					-	0.666	-	-
	epochs	**context**	**features**	**threshold**				
`ModernGBERT_1B`	0	none	none	0.042	0.2958	-	0.6366	-
`ModernGBERT_1B`	0	sentence	none	0.042	0.3341	-	0.5081	-
`ModernGBERT_1B`	0	sentence	all	0.042	0.3947	-	0.2667	-
`ModernGBERT_1B`	5	none	none	0.042	-	0.7626	-	0.0533
`ModernGBERT_1B`	5	sentence	none	0.042	-	0.7921	-	0.0499
`ModernGBERT_1B`	5	sentence	all	0.042	-	0.7829	-	0.0505
`LLaMmlein2Vec_7B`	0	none	none	0.042	0.2976	-	0.4464	-
`LLaMmlein2Vec_7B`	0	sentence	none	0.042	0.2958	-	0.8632	-
`LLaMmlein2Vec_7B`	0	sentence	all	0.042	0.2958	-	0.7799	-
`LLaMmlein2Vec_7B`	5	none	none	0.042	-	0.7633	-	0.0528
`LLaMmlein2Vec_7B`	5	sentence	none	0.042	-	**0.7934**	-	**0.0485**
`LLaMmlein2Vec_7B`	5	sentence	all	0.042	-	0.7921	-	0.0492
`LLaMmlein_7B`	0	none	none	0.042	0.3383	-	0.3332	-
`LLaMmlein_7B`	0	sentence	none	0.042	0.2958	-	0.6449	-
`LLaMmlein_7B`	0	sentence	all	0.042	0.2958	-	0.6952	-
`LLaMmlein_7B`	5	none	none	0.042	-	0.7808	-	0.0523
`LLaMmlein_7B`	5	sentence	none	0.042	-	0.7931	-	0.0490
`LLaMmlein_7B`	5	sentence	all	0.042	-	0.7927	-	0.0488

Table 3: Binary ($F1$) and probabilistic classification (MAE) results. Our results are rounded to the fourth digit after the floating point. Prior best results before are <u>underlined</u> and current best results are in **bold**. All reported binary scores use the selected threshold $\tau = 0.042$ (Section 3.4).

outputs at $\tau = 0.042$ selected on the development set (Section 3.4). We first compare our models to previously reported baselines and then analyze the impact of context and feature injection.

4.1. Comparison to Prior Work

Among previously reported systems, the original shared-task baseline Yimam et al. (2018) achieves the strongest binary performance on the test set ($F1_{macro} = 0.7546$), while TMU (Kajiwara and Komachi, 2018) reports the best probabilistic performance ($MAE = 0.0610$). Later LLM-based approaches evaluated by Smădu et al. (2024) reach test $F1_{macro}$ scores between 0.666 and 0.708, remaining below the strongest feature-based baselines for German.

Across all three model families evaluated in this work, fine-tuned configurations substantially outperform both prior LLM-based systems and earlier feature-based baselines. The best binary performance on the test set is achieved by the contextualized `LLaMmlein2Vec_7B` model with $F1_{macro} = 0.7934$, followed closely by `LLaMmlein_7B` with $F1_{macro} = 0.7931$. For probabilistic CWI, the lowest MAE is 0.0485 (LLaMmlein2Vec_7B, contextualized), improving considerably over the previously best reported MAE of 0.0610. To the best of our knowledge, these results constitute the strongest reported performance on the German CWIG3G2 benchmark under directly comparable evaluation settings.

4.2. Pretrained-only vs. Fine-Tuning

We use pretrained-only to denote evaluation of the pretrained sequence-classification models without any task-specific fine-tuning (0 training epochs);

this is not a prompting-based setup. We report these pretrained-only results on the development set to characterize the starting point before fine-tuning but do not report on the test set to avoid bias the experimental decision process based on test results. For instance, LLaMmlein_7B achieves $F1_{macro} = 0.3383$ (no context) and 0.2958 (with context) on the development set, with corresponding MAE values between 0.3332 and 0.6952 depending on the configuration. Similar patterns are observed for LLaMmlein2Vec_7B and ModernGBERT_1B. Fine-tuning dramatically reduces MAE and improves macro-F_1 across all model families, moving from weak pretrained-only development performance to strong test performance after five epochs.

4.3. Effect of Sentence-Level Context (RQ1)

To address RQ1, we compare target-only configurations with contextualized setups. For both LLaMA-style models, incorporating sentence context consistently improves performance.

For `LLaMmlein2Vec_7B`, adding context increases test $F1_{macro}$ from 0.7633 (no context) to 0.7934 (sentence context), while reducing MAE from 0.0528 to 0.0485. A similar trend holds for `LLaMmlein_7B`, where contextualization improves $F1_{macro}$ from 0.7808 to 0.7931 and reduces MAE from 0.0523 to 0.0490.

`ModernGBERT_1B` shows the same qualitative behavior: contextualization increases test $F1_{macro}$ from 0.7626 to 0.7921 and reduces MAE from 0.0533 to 0.0499.

Overall, sentence-level context yields consistent gains across architectures and evaluation metrics, supporting RQ1: contextual information positively affects German CWI performance in fine-tuned LLMs.

4.4. Effect of Feature Injection (RQ2)

To address RQ2, we compare contextualized setups with and without injected statistical features. The results do not indicate consistent improvements from feature injection.

For `LLaMmlein2Vec_7B`, adding features slightly decreases binary performance ($0.7934 \rightarrow 0.7921$) and slightly increases MAE ($0.0485 \rightarrow 0.0492$). For `LLaMmlein_7B`, feature injection leads to a small reduction in $F1_{macro}$ ($0.7931 \rightarrow 0.7926$) with only marginal changes in MAE ($0.0490 \rightarrow 0.0488$).

For `ModernGBERT_1B`, the differences between contextualized configurations with and without features are small and do not reveal a systematic pattern. Depending on the threshold and metric, feature injection may result in slight improvements or slight degradations, but effect sizes remain minor.

On the test set, feature injection does not yield consistent gains over the contextualized setup without features; differences are small and sometimes negative. On the development set, feature injection can improve macro-F_1 in some configurations, suggesting that its benefit may be unstable and sensitive to split/model interactions.

4.5. Threshold Sensitivity (RQ3)

RQ3 investigates how sensitive binary CWI performance is to the choice of decision threshold when models are trained on probabilistic labels. Since our models output continuous scores $p \in [0, 1]$, binary predictions require selecting a threshold τ.

For the fine-tuned models (5 epochs), development performance is highest at $\tau = 0.042$ across all model families and setups, and it decreases for larger thresholds ($\tau = 0.083$, $\tau = 0.167$). For example, for `LLaMmlein2Vec_7B` (5 epochs, contextualized), dev $F1_{macro}$ decreases from 0.8170 at $\tau = 0.042$ to 0.7862 at $\tau = 0.083$ and 0.6187 at $\tau = 0.167$. In contrast, pretrained-only (0 epochs) results do not follow a consistent monotonic pattern with respect to τ, which is expected because the unfitted classifier head yields poorly calibrated scores.

Importantly, although the gold probabilistic labels are quantized in steps of approximately $1/12$, the decision threshold still meaningfully affects predictions derived from model outputs. Selecting τ based on development performance therefore remains necessary. In our experiments, $\tau = 0.042$ consistently yields the highest development performance and is fixed for all reported test results.

Overall, binary performance is sensitive to the threshold choice, but the ranking between model architectures and experimental setups remains stable across the tested values. This indicates that while absolute $F1_{macro}$ scores vary with τ, the qualitative conclusions of this study are robust to reasonable threshold variations.

5. Conclusion

In this work, we evaluated state-of-the-art German-pretrained LLMs on complex word identification and compared them to established feature-based baselines and previously reported LLM systems. Across all three model families, fine-tuned configurations substantially outperform pretrained-only setups and set new benchmarks for both binary ($F1_{macro}$) and probabilistic (MAE) German CWI.

Our experiments show that sentence-level context consistently improves performance, confirming that lexical complexity is not purely a property

of isolated target expressions but benefits from contextual modeling. In contrast, injecting explicit length- and frequency-based features does not yield consistent gains once models are fine-tuned, suggesting that such information may already be implicitly encoded in pretrained representations.

These findings indicate that German CWI can be effectively addressed using compact, parameter-efficient fine-tuning of modern LLMs. At the same time, the large performance gap between pretrained-only and fine-tuned settings highlights that lexical complexity prediction remains a task requiring targeted supervision. Finally, we emphasize that complexity judgments are inherently audience-dependent (Gooding et al., 2021), and future work should further investigate personalized and domain-specific modeling approaches for German CWI.

6. Outlook

Beyond complex word identification as a standalone task, our models can be interpreted as a proxy for lexical complexity, thereby serving both as a modeling signal and as an evaluation bridge for LCP. In this sense, CWI approximates a decision boundary in the underlying complexity space, complementing continuous LCP approaches that aim to model graded difficulty.

A further use case is supporting the evaluation of lexical complexity models. In particular, token- or span-level complexity estimates can be aggregated to text-level signals when combined with complementary indicators (e.g., sentence-level properties Schomacker et al., 2024). Such composite assessments may provide a more informative view of text complexity than relying solely on classical readability formulas (Tanprasert and Kauchak, 2021). Initial findings (e.g., Deilen et al., 2023; Anschütz et al., 2023; Schomacker et al., 2026) showcase LLMs used end-to-end without necessarily relying on CWI as "classic" pipelines did. Having CWI-specific models still offers benefits in terms interpretability and modularity.

Although newer models will likely be released in the future, our results suggest that the qualitative trends observed in this study may extend beyond the specific models evaluated here. In particular, incorporating sentence-level context was associated with improved performance in our experiments, whereas explicit feature injection did not consistently lead to additional gains under the configurations we tested.

Finally, the BERT-based model (`ModernGBERT_1B`) may represent a practical option for deployment scenarios. Despite having substantially fewer parameters than the 7B LLaMA-style models, it achieves comparable performance in our experiments. This indicates that smaller models can, under certain conditions, provide a reasonable balance between computational efficiency and predictive quality.

7. Limitations

First, our models are trained to predict probabilistic complexity scores and we derive binary predictions by thresholding the predicted score at a fixed decision threshold ($\tau = 0.042$), selected on the development set. While this avoids tuning on the test set, binary $F1_{macro}$ values are nevertheless dependent on the chosen decision rule and should be interpreted together with the threshold sensitivity results (RQ3). In addition, the probabilistic gold labels in CWIG3G2 are quantized in steps of approximately $1/12$ (smallest non-zero value 0.083), which constrains how finely label semantics can be reflected in binary decision boundaries.

Second, the dataset is relatively small (7905 instances across train/dev/test in our split) and restricted to the CWIG3G2 domain (news) and annotation population. It nevertheless remains the only publicly available German word-level CWI dataset. We therefore prioritize controlled comparison over large-scale pretraining variation. As a result, the reported improvements may not fully generalize to other German genres, domains (e.g., administrative texts, health communication), or reader groups with different proficiency profiles.

Third, our feature injection approach is intentionally simple and may not fully exploit structured scalar features. Therefore, our findings should not be interpreted as evidence that lexical features are generally unhelpful, but rather that this particular integration strategy does not provide additional benefit over contextualized transformer representations. Our approach may underutilize the features and is sensitive to tokenization and formatting choices. More structured integration methods, e.g., dedicated feature embeddings, could yield different outcomes and remain for future work.

Finally, we evaluate only three pretrained model families and a single compute setting (one GPU type) with a fixed hyperparameter configuration. Although we align our setup with prior work for comparability, alternative choices (e.g., different LoRA target modules, ranks, quantization settings, or longer training) might affect absolute performance and potentially interact with the impact of context and feature injection. In all experiments a single random seed was used, following prior benchmark setups. While this facilitates comparability, we acknowledge that reporting variance across multiple seeds would further strengthen robustness claims.

Ethics statement

This work explores large language models for complex word identification (CWI) in German, with the goal of supporting accessibility and text simplification. While the task itself poses low direct risk, several ethical considerations apply.

Pretrained language models may encode societal biases present in their training data, which could affect how lexical complexity is assessed. In addition, the notion of "complexity" is inherently subjective and depends on reader characteristics such as language proficiency and background knowledge. Our models are trained on annotated data that reflect specific annotator perspectives and may not generalize to all user groups.

CWI systems should therefore be used as supportive tools rather than replacements for human judgment, particularly in sensitive contexts. Careful evaluation is required before deployment to ensure fair and appropriate use.

Lay Summary

Some words are harder to understand than others. This can make texts difficult to read, especially for people with lower reading skills or those learning a language. Our research looks at how artificial intelligence (AI) can automatically find these difficult words. This task is called complex word identification (CWI).

We focus on German and test modern AI language models to see how well they can recognize difficult words. These models are already trained on large amounts of text, but we also adapt them further using a small amount of task-specific data. This process is called fine-tuning.

Our results show that these fine-tuned models can identify difficult words very accurately. They perform better than earlier methods that rely heavily on manually designed features, such as word length or how often a word appears in texts.

We also find that context matters: a word may be easy or difficult depending on the sentence it appears in. Models that consider the full sentence perform better than those that look at words in isolation.

Overall, our work shows that modern AI models can be an effective tool for detecting difficult words in German sentences. This could support applications such as simplifying texts, improving accessibility, or helping people better understand written information.

8. Bibliographical References

Miriam Anschütz, Joshua Oehms, Thomas Wimmer, Bart\lomiej Jezierski, and Georg Groh. 2023. Language Models for German Text Simplification: Overcoming Parallel Data Scarcity through Style-specific Pre-training. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1147–1158, Toronto, Canada. Association for Computational Linguistics. URL: https://aclanthology.org/2023.findings-acl.74, doi:10.18653/v1/2023.findings-acl.74.

Joachim Bingel and Johannes Bjerva. 2018. Cross-lingual complex word identification with multitask learning. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 166–174, New Orleans, Louisiana, USA. Association for Computational Linguistics. URL: https://aclanthology.org/W18-0518/, doi:10.18653/v1/W18-0518.

Silvana Deilen, Sergio Hernández Garrido, Ekaterina Lapshinova-Koltunski, and Christiane Maaß. 2023. Using ChatGPT as a CAT tool in Easy Language translation. In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 1–10, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria. URL: https://aclanthology.org/2023.tsar-1.1/.

Pierre Finnimore, Elisabeth Fritzsch, Daniel King, Alison Sneyd, Aneeq Ur Rehman, Fernando Alva-Manchego, and Andreas Vlachos. 2019. Strong Baselines for Complex Word Identification across Multiple Languages. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 970–977, Minneapolis, Minnesota, USA. Association for Computational Linguistics. URL: https://aclanthology.org/N19-1102, doi:10.18653/v1/N19-1102.

Sian Gooding. 2023. *A Personalised Approach to Lexical Complexity*. Doctoral Thesis, University of Cambridge, Cambridge, England. URL: https://doi.org/10.17863/CAM.116567.

Sian Gooding and Ekaterina Kochmar. 2019. Complex Word Identification as a Sequence Labelling Task. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1148–1153, Florence, Italy. Association for Computational Linguistics. URL: https://aclanthology.org/P19-1109/, doi:10.18653/v1/P19-1109.

Sian Gooding, Ekaterina Kochmar, Seid Muhie Yimam, and Chris Biemann. 2021. Word Complexity is in the Eye of the Beholder. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4439–4449, Online. Association for Computational Linguistics. URL: https://aclanthology.org/2021.naacl-main.351, doi:10.18653/v1/2021.naacl-main.351.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *Proceedings of the International Conference on Learning Representations*, Online. URL: https://openreview.net/forum?id=nZeVKeeFYf9.

Tomoyuki Kajiwara and Mamoru Komachi. 2018. Complex Word Identification Based on Frequency in a Learner Corpus. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 195–199, New Orleans, Louisiana, USA. Association for Computational Linguistics. URL: https://aclanthology.org/W18-0521/, doi:10.18653/v1/W18-0521.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Re, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic Evaluation of Language Models. *Transactions on Machine Learning Research*. URL: https://openreview.net/forum?id=iO4LZibEqW.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, Benjamin Bossan, and Marian Tietz. 2022. PEFT: State-of-the-art parameter-efficient fine-tuning methods. URL: https://github.com/huggingface/peft.

Tomoya Mizumoto, Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. 2011. Mining Revision Log of Language Learning SNS for Automated Japanese Error Correction of Second Language Learners. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 147–155, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Babak Naderi, Salar Mohtaj, Kaspar Ensikat, and Sebastian Möller. 2019. Subjective Assessment of Text Complexity: A Dataset for German Language. Version Number: 1. URL: https://arxiv.org/abs/1904.07733, doi:10.48550/ARXIV.1904.07733.

Jan Pfister and Andreas Hotho. 2024. SuperGLEBer: German Language Understanding Evaluation Benchmark. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7904–7923, Mexico City, Mexico. Association for Computational Linguistics. URL: https://aclanthology.org/2024.naacl-long.438/, doi:10.18653/v1/2024.naacl-long.438.

Jan Pfister, Julia Wunderle, and Andreas Hotho. 2025. LLäMmlein: Transparent, Compact and Competitive German-Only Language Models from Scratch. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2227–2246, Vienna, Austria. Association for Computational Linguistics. URL: https://aclanthology.org/2025.acl-long.111/, doi:10.18653/v1/2025.acl-long.111.

Thorben Schomacker, Miriam Anschütz, Regina Stodden, Georg Groh, and Marina Tropmann-Frick. 2024. Overview of the GermEval 2024 Shared Task on Statement Segmentation in German Easy Language (StaGE). In *Proceedings of GermEval 2024 Shared Task on Statement Segmentation in German Easy Language (StaGE)*, pages 1–14, Vienna, Austria. Association for Computational Lingustics. URL: https://aclanthology.org/2024.germeval-1.1.

Thorben Schomacker, Burak Tinman, Chris Biemann, and Marina Tropmann-Frick. 2026. LLMs for Easy Language Translation: A Case Study on German Public Authorities Web Pages. In *KI 2025: Advances in Artificial Intelligence*, pages 252–261, Cham. Springer Nature Switzerland. doi:10.1007/978-3-032-02813-6_20.

Laura Seiffe, Fares Kallel, Sebastian Möller, Babak Naderi, and Roland Roller. 2022. Subjective Text Complexity Assessment for German. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 707–714, Marseille, France. European Language Resources

Association. URL: https://aclanthology.org/2022.lrec-1.74.

Matthew Shardlow. 2013. A Comparison of Techniques to Automatically Identify Complex Words. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pages 103–109, Sofia, Bulgaria. Association for Computational Linguistics. URL: https://aclanthology.org/P13-3015/.

Matthew Shardlow. 2015. *Lexical Simplification: Optimising the Pipeline*. Dissertation, University of Manchester, Manchester, United Kingdom. URL: https://pure.manchester.ac.uk/ws/portalfiles/portal/54575176/FULL_TEXT.PDF.

Matthew Shardlow, Fernando Alva-Manchego, Riza Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, Anna Hülsing, Yusuke Ide, Joseph Marvin Imperial, Adam Nohejl, Kai North, Laura Occhipinti, Nelson Peréz Rojas, Nishat Raihan, Tharindu Ranasinghe, Martin Solis Salazar, Sanja Štajner, Marcos Zampieri, and Horacio Saggion. 2024. The BEA 2024 Shared Task on the Multilingual Lexical Simplification Pipeline. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 571–589, Mexico City, Mexico. Association for Computational Linguistics.

Răzvan-Alexandru Smădu, David-Gabriel Ion, Dumitru-Clementin Cercel, Florin Pop, and Mihaela-Claudia Cercel. 2024. Investigating Large Language Models for Complex Word Identification in Multilingual and Multidomain Setups. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16764–16800, Miami, Florida, USA. Association for Computational Linguistics. URL: https://aclanthology.org/2024.emnlp-main.933/, doi:10.18653/v1/2024.emnlp-main.933.

Teerapaun Tanprasert and David Kauchak. 2021. Flesch-Kincaid is Not a Text Simplification Evaluation Metric. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 1–14, Online. Association for Computational Linguistics. URL: https://aclanthology.org/2021.gem-1.1, doi:10.18653/v1/2021.gem-1.1.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2019/hash/4496bf24afe7fab6f046bf4923da8de6-Abstract.html.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics. URL: http://aclweb.org/anthology/W18-5446, doi:10.18653/v1/W18-5446.

Julia Wunderle, Anton Ehrmanntraut, Jan Pfister, Fotis Jannidis, and Andreas Hotho. 2025. New Encoders for German Trained from Scratch: Comparing ModernGBERT with Converted LLM2Vec Models. ArXiv:2505.13136 [cs]. URL: http://arxiv.org/abs/2505.13136, doi:10.48550/arXiv.2505.13136.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. A Report on the Complex Word Identification Shared Task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 66–78, New Orleans, Louisiana, USA. Association for Computational Linguistics. URL: https://aclanthology.org/W18-0507/, doi:10.18653/v1/W18-0507.

Seid Muhie Yimam, Sanja Štajner, Martin Riedl, and Chris Biemann. 2017. Multilingual and Cross-Lingual Complex Word Identification. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 813–822, Varna, Bulgaria. INCOMA Ltd. URL: https://aclanthology.org/R17-1104/, doi:10.26615/978-954-452-049-6_104.

Book Complexity Level Assignment in French and Portuguese

Jorge Baptista[1,2], **David Antunes**[2], **Wafa Aissa**[3], **Julien Zakhia Doueihi**[3]
Trang Pham Tran Hanh[3], **Eugénio Ribeiro**[2,4], **Thomas François**[3], **Raquel Amaro**[5]
[1]U. Algarve, Portugal; [2]INESC-ID Lisboa, Portugal; [3]UCLouvain, Belgium;
[4]Iscte-IUL, Portugal; [5]U. NOVA Lisboa, Portugal
{jorge.baptista, david.f.l.antunes, eugenio.ribeiro}@inesc-id.pt;
{wafa.aissa, julien.zakhia, tran.pham, thomas.francois}@uclouvain.be;
raquelamaro@fcsh.unl.pt

Abstract

Selecting reading materials suitable for adults with low literacy levels is a challenging task in Adult Learning (AL) contexts, particularly when dealing with full books or long texts, as textual complexity may be difficult to infer from the analysis of a small set of extracted samples. This paper presents experiments on estimating the complexity level of full-length books in Portuguese and French, with the aim of identifying the most effective sampling and result aggregation procedure. The procedure is supported by a natural language processing and machine-learning-based system for the automatic assignment of textual complexity levels. Complexity levels follow a scale specifically devised for AL audience, aligned with internationally recognised scales, ranging from L1 (*Very Easy*) to L4 (*More Complex*). The ultimate goal is to guide the expansion of a publicly accessible database of book titles with reliable complexity information, particularly benefiting key stakeholders in AL, such as students and teachers, as well as librarians and publishers concerned with literacy promotion.

Keywords: Text Complexity Assignment, Full Books, Sampling, Adult Learning, French, Portuguese

1. Introduction and Motivation

Selecting reading materials that are appropriate for adults with low literacy skills remains a central challenge in Adult Learning (AL) contexts. This challenge becomes particularly acute when the unit of analysis is not a short passage but a full book or long-form text, where internal heterogeneity in lexical, syntactic, and discourse-level properties makes global readability estimation non-trivial (Li et al., 2024). In practice, librarians, educators, and publishers often need to make decisions about the suitability of books for specific learner populations without access to complete texts, relying instead on partial excerpts or limited samples, and personal intuition.

Within the iRead4Skills project[1] (iR4S), textual complexity is operationalised through a four-level scale (L1–L4) specifically designed to address low-literacy adult learners' needs (Amaro et al., 2025; Monteiro et al., 2023), and aligned with CEFR and PIACC-inspired notions of linguistic difficulty (Council of Europe, 2020; OECD, 2013a,b, 2021). While CEFR primarily targets second-language learners, its descriptors have been adapted in iR4S to address functional literacy in adult native populations. Nevertheless, the extent to which these levels generalise across different populations remains an open question.

Whereas classic readability research explicitly addressed sampling procedures for long texts, most recent natural language processing (NLP)-based approaches demonstrate strong performance at the short-text or passage level (Collins-Thompson and Callan, 2005; Ribeiro et al., 2024b,a). The methodological and empirical implications of extrapolating such passage-level models to full-book assessment, however, remain comparatively underexplored. A naive aggregation of passage-level predictions can lead to biased or unstable estimates, particularly for long texts that exhibit substantial variation in difficulty across chapters or sections.

While recent large language models can process longer contexts, their effective input length remains constrained relative to full-length books, particularly for texts exceeding tens of thousands of tokens. Moreover, such approaches entail high computational costs and reduced interpretability when applied in operational settings such as Adult Learning. In contrast, sampling-based strategies offer a lightweight, transparent, and controllable alternative, compatible with existing readability models and realistic deployment constraints (e.g., partial access to texts, limited processing resources).

This paper addresses this gap by investigating principled strategies for estimating the global complexity level of full-length books in Portuguese and French based on a limited number of textual samples and on the iR4S complexity analysis (see Appendix C). Our objective is twofold: first, we aim to identify sampling and aggregation strategies that yield robust and interpretable book-level complexity estimates under realistic usage constraints, such as partial text availability and limited compu-

[1] https://iread4skills.com/

Proceedings of the Joint Workshop on Readability and Text Simplification (READIxTSAR) @ LREC 2026, pages 12–25
11 May 2026.

tational resources; second, we seek to empirically validate these strategies against human-annotated gold standards, on the one hand, and against full-text analysis, on the other, assessing their reliability and practical adequacy for deployment in AL and literacy promotion scenarios.

The work reported here builds directly on earlier methodological discussions within iR4S concerning lightweight yet informed sampling protocols for long texts, and is explicitly inspired by recent advances in long-document readability modelling. In particular, we adapt the conceptual framework proposed by Li et al. (2024), which treats book-level readability as an emergent property arising from the distribution of difficulty across multiple textual segments rather than from isolated passages. By grounding our experiments in this perspective, we aim to contribute empirical evidence and concrete methodological guidance for book-level readability assessment in diversified settings.

2. Related Work: Readability Beyond Passages

Automatic readability assessment has a long tradition in educational research and Natural Language Processing. Early approaches relied on handcrafted readability formulas—such as Flesch (Flesch, 1948), Dale–Chall (Dale and Chall, 1948; Chall and Dale, 1995), and SMOG (McLaughlin, 1969)—which operationalised textual difficulty through a small set of shallow surface-level proxies, including sentence length, syllable-based word length, or predefined lexical lists. While these formulas remain influential in educational practice, particularly in institutional and pedagogical settings, their limited linguistic coverage, reliance on linear models, and reduced sensitivity to discourse, syntactic, and semantic variation restrict their applicability across genres, populations, and languages (Benjamin, 2012; Collins-Thompson, 2014).

These limitations are especially salient in Adult Learning (AL) contexts, where textual complexity must align with functional literacy goals and proficiency scales, such as CEFR. As a result, research has progressively shifted towards data-driven, non-linear machine learning approaches (Feng et al., 2010), capable of modelling richer lexical, syntactic, and semantic representations of text difficulty.

Subsequent work reframed readability assessment as a supervised classification or regression task, exploiting a wider range of linguistically-informed features. These include lexical frequency and diversity measures, morphosyntactic complexity indicators, and discourse-level cues, often combined within statistical or kernel-based classifiers (Schwarm and Ostendorf, 2005; Petersen and Ostendorf, 2009; Sheehan et al., 2010; François and Fairon, 2012). More recent studies have incorporated neural representations, either as standalone predictors or in hybrid architectures that integrate pre-trained embeddings with handcrafted linguistic features (Deutsch et al., 2020; Lee et al., 2021; Wilkens et al., 2024). These approaches consistently report strong performance at the passage or short-text level.

However, the vast majority of readability research remains fundamentally passage-oriented. Large-scale readability corpora typically consist of short, decontextualised excerpts, carefully controlled for length and often sampled from pedagogical materials (Crossley et al., 2023). While such design choices are methodologically sound for passage-level prediction, they implicitly assume textual homogeneity and fail to account for the internal variation that characterises long documents and full books.

The limitations of passage-based approaches for long texts have long been acknowledged in applied settings. Educational and psychometric guidelines already recommend sampling multiple excerpts from different parts of a book and aggregating their scores to improve reliability (Allan et al., 2005; Stenner et al., 2006). These practices recognise positional effects and within-text variation, but they predate contemporary NLP models and lack empirical validation under modern computational paradigms.

Recent advances in long-document modelling have focused primarily on architectural solutions for processing long sequences, such as sparse-attention transformers and sequence compression mechanisms (Beltagy et al., 2020; Zaheer et al., 2020). While these models extend the maximum input length, they do not directly address the conceptual mismatch between passage-level difficulty signals and book-level readability, nor do they encode explicit notions of difficulty distribution within a text.

A decisive step towards bridging this gap is provided by Li et al. (2024), who explicitly argue that book-level readability cannot be reliably inferred from isolated passages or naive truncation strategies. Through extensive experiments, they demonstrate that direct transfer from passage-level models leads to systematic bias, typically overestimating difficulty for long texts. Their work introduces the notion of books as compositions of multiple "difficulty fragments", whose distribution and aggregation are informative for global readability estimation. Crucially, they show that aggregation strategies and sampling density play a more decisive role than raw model capacity when moving from passages to books.

The present work adopts this perspective while departing from purely model-centric solutions.

Rather than proposing a new end-to-end neural architecture for long texts, we focus on sampling-aware and aggregation-aware strategies that are compatible with realistic usage scenarios, namely those involving the use of the iR4S system by trainers in AL contexts, as well as by librarians and publishers with limited time, to classify books according to their level of complexity. In particular, we build on established multi-sample readability practices (Allan et al., 2005; Stenner et al., 2006), recent corpus-based insights on controlled excerpt selection (Crossley et al., 2023), and the distributional view of difficulty advanced by Li et al. (2024). This allows us to investigate how lightweight sampling regimes and simple aggregation operators can yield robust and interpretable book-level complexity estimates across languages, even under constrained access to full textual data.

3. Automatic Book-Level Complexity Estimation

3.1. Corpus and Sampling Procedure

Two corpora of full-length texts were compiled for the experiments, one per language. The Portuguese corpus was constructed following a two-step procedure. The iR4S corpus (Pintard et al., 2024) contains short excerpts from longer texts that had previously been assigned a complexity level. In order to enable full-text analysis, the complete digital versions of the corresponding source texts (e.g., full books or documents) were retrieved for each excerpt and incorporated into the new corpus.

Second, additional texts from publicly available repositories were included to ensure balance across complexity levels. The overall classification of the texts was validated by experts familiar with the iR4S classification framework (Levels L1–L4) and specifically trained for this level-assignment task. The final corpus comprises 16 texts, with four texts per level, totalling 943,169 words. Corpus details are provided in Appendix A.

For the French corpus, we initially retrieved open-access books from the ABU repository[2], but preliminary experiments revealed that such corpus includes nearly only books from the 'More Complex' level (L4). As we need a fair representation of the four levels of the scale used, we then selected 29 simplified readers published for French as a foreign language. These books are either simplified version of classic literary work (e.g., *Carmen*, *Germinal*) or original stories adapted for learners. According to their CEFR levels, these texts can be classified as L1 ('Very Easy'), roughly corresponding to A1, L2 ('Easy') roughly corresponding to A2, and L3 ('Plain'), roughly corresponding to B2. (cf.Table 7 in Appendix B for their name and other details). Together with the previously collected titles, the compiled corpus covers all iR4S complexity levels. Selected books were scanned with optical character recognition tools and manually revised. Both corpora consist of full-text books classified according to the iR4S levels. Due to copyright constraints, full texts of the corpora cannot be distributed. However, sampling procedures, model descriptions, and aggregated annotations are documented in sufficient detail in Appendices A and B to ensure methodological reproducibility.

While the two corpora (Portuguese and French) are not strictly parallel in composition, particularly due to the inclusion of simplified readers in French to ensure level balance, both datasets are aligned with the same iR4S complexity scale and validated by expert annotation. The goal of the study is not strict cross-lingual comparison but rather the evaluation of sampling and aggregation strategies across two typologically different yet methodologically compatible settings.

All texts were manually cleaned to remove paratextual material (e.g., headers, footers, tables, figures and captions, indices, footnotes, and references).

A dedicated sampling tool was developed to automatically segment text files into excerpts of approximately equal length. The tool allows fine-grained control over the minimum and maximum size of each excerpt, while preserving paragraph boundaries and preventing paragraph splits. It also enables users to specify the number of excerpts to be extracted. The selected excerpts are drawn from different sections of the text and are evenly distributed throughout the document.

For the purposes of our experiments, the sampling tool first segments each text into as many valid excerpts as possible. It then selects a fixed set of 10 samples per text. These samples are drawn from evenly distributed *loci* within the book (beginning, 10%, 20%, . . . , end). Each sample contains between 250 and 300 words (approximately one printed page) and preserves paragraph boundaries. Ideally, all selected texts would have had a minimum length of 5,000 words. Two texts in the Portuguese corpus were shorter; from these, only 2 and 8 samples, respectively, could be extracted in compliance with the constraints described above.

3.2. Experiment 1: Sampling-Based Automatic Estimation

The first experiment examines the estimation of overall book complexity based exclusively on automatically predicted complexity levels assigned to sampled text segments. It is conducted for both Portuguese and French and reflects a realistic de-

[2] http://abu.cnam.fr/

ployment scenario in which no manual expert annotations are available for the full books.

Each book is segmented into fixed-length excerpts according to the sampling protocol described in Subsection 3.1. Segment-level complexity is then automatically estimated using the iR4S complexity assessment model for each language (see Appendix C), resulting in a distribution of predicted complexity levels across the book.

To derive a single book-level complexity estimate, we evaluate several aggregation strategies, including measures of central tendency (mean and median), extremal values (minimum and maximum), frequency-based aggregation, and quantile-based approaches (25th and 75th percentiles).

The primary objective of this experiment is to assess how well a reduced number of samples or excerpts approximates the complexity profile obtained from a denser representation of the book, based on *all* excerpts into which it can be segmented.

Reduced sampling configurations (ranging from 1 to 10 excerpts) are then compared against this reference profile using correlation and error-based metrics, including Mean Absolute Error (MAE) and Spearman's rank correlation coefficient. This allows us to quantify how closely the reduced samples approximate the estimated full-book complexity level. For each sampling configuration, excerpts are selected from evenly distributed positions throughout the text (e.g., $S = 5$ corresponds to samples drawn at 20%, 40%, 60%, 80%, and 100% of the text length).

Table 1 presents the results of this experiment for Portuguese texts, while Table 2 reports the corresponding results for French texts.

For the Portuguese books, Table 1 presents the approximation performance of reduced excerpt samples across aggregation strategies and sample sizes, while Figure 1 in Appendix C illustrates the corresponding trends. Overall, performance improves as the number of sampled excerpts (S) increases, although the progression is not strictly monotonic across all aggregation strategies.

Overall, MAE decreases with increasing S, although convergence behaviour differs across aggregation strategies. Central-tendency estimators (mean, median, mode) stabilise more quickly, yielding consistently low error from moderate sampling levels onward ($S \geq 5$). In particular, the mean reaches minimal error at $S = 7$ (MAE = 0.0000) and remains stable thereafter. In contrast, extreme-value strategies (min, max) exhibit higher variability and less consistent improvement, while quantile estimators (Quantile (Q25), Quantile (Q75)) show intermediate behaviour, with higher error at small S followed by gradual stabilisation.

The Spearman's ρ results mirror these patterns. Mean and median aggregation maintain relatively high and stable rank correlations across sampling levels, indicating preserved book ordering under reduced sampling. Quantile estimators improve steadily with larger S, whereas extreme-value methods display greater fluctuation in ranking consistency.

Paired t-tests on per-book differences confirm these patterns. Central-tendency estimators show no significant deviation from the baseline at any sampling level (mean: $p = 0.43$ at $S = 1$, $p = 0.97$ at $S = 6$, all $p > 0.16$ overall; median and mode similarly non-significant across all S). In contrast, the maximum aggregator is significant for all sampling levels ($S = 1$–10; all $p \leq 0.041$), and the minimum is significant for every $S \neq 8$ (all $p < 0.05$). Quantile estimators exhibit limited early instability: Q25 is significant at $S = 1$ ($p = 0.029$) but not thereafter, while Q75 is significant at $S = 2$ ($p = 0.009$) and non-significant for larger S.

Taken together, the results demonstrate that central-tendency aggregation is highly robust to reduced sampling in Portuguese. Moderate sampling levels ($\approx$ 6–8) are sufficient to achieve low error, stable ranking, and no statistically significant deviation from the full-book baseline, whereas extreme-value methods remain comparatively unstable and systematically biased.

For the French books, Table 2 reports the book-difficulty approximation performance of reduced excerpt samples across aggregation strategies and sample sizes (Figure 2 illustrates the corresponding trends). We observe that performance improves consistently as the number of sampled excerpts (S) increases. For central-tendency aggregators, MAE decreases markedly between $S = 1$ and $S = 5$ and stabilises from $S \geq 7$, while Spearman's ρ exceeds 0.95 from $S = 7$ onward, indicating strong ranking consistency with the baseline.

Paired t-tests on per-book differences were conducted to assess whether reduced sampling introduces systematic deviation from the full-excerpt baseline. Results reveal a sample-size effect for quantile and extreme-value aggregators, whereas central-tendency estimators (mean, median, mode) remain stable across sampling levels. Mean aggregation shows no significant deviation at any S (e.g., $p = 0.43$ at $S = 1$, $p = 0.97$ at $S = 6$), and similar patterns are observed for median and mode. In contrast, Q25 is significant for $S = 1$–3 ($p = 0.0006$ at $S = 1$) before stabilising, while Q75 remains significant up to $S = 5$ ($p = 0.0040$ at $S = 1$). Extreme-value methods are the least stable: min is significant from $S = 1$ to $S = 9$ ($p < 0.01$), and max is significant for $S = 1$–7 and again at $S = 10$ ($p = 0.0433$). Overall, increasing S reduces systematic deviation for robust aggregation strategies, but extreme-value approaches remain persistently biased.

Table 1: Approximation of Portuguese full-book complexity using reduced samples. S denotes the number of samples. Performance is measured using Mean Absolute Error (MAE ↓) and Spearman's rank correlation coefficient (ρ ↑). Best values in each column are marked in bold. Best values in each row are marked in italic.

	Mean		Max		Min		Median		Mode		Q25		Q75	
S	MAE ↓	ρ ↑	MAE ↓	ρ ↑	MAE ↓	ρ ↑	MAE ↓	ρ ↑	MAE ↓	ρ ↑	MAE ↓	ρ ↑	MAE ↓	ρ ↑
1	*0.3125*	*0.6873*	0.7500	0.4927	0.8125	0.4849	0.3750	0.6407	0.4375	0.5526	0.5625	0.4277	*0.3125*	*0.6873*
2	0.5000	0.5032	0.5625	*0.7037*	0.5000	0.6551	0.3125	0.6184	*0.2500*	0.7032	*0.2500*	0.6990	0.3750	*0.7037*
3	0.3125	0.6407	0.5000	0.7062	0.4375	0.5952	0.3125	0.6793	0.2500	0.7868	0.3125	0.6587	*0.1250*	*0.8615*
4	*0.1250*	0.8281	0.4375	0.7174	0.3750	0.6708	*0.1250*	0.8502	0.2500	0.6859	*0.1250*	*0.8783*	0.2500	0.7075
5	*0.1250*	0.7570	0.5625	0.7037	0.3750	0.6708	0.2500	0.6762	0.2500	0.7032	0.1250	0.8783	*0.1250*	*0.9074*
6	0.3125	0.6407	0.4375	0.7174	0.3750	0.6708	0.3125	0.6793	0.2500	0.7868	0.3125	0.5928	*0.1250*	*0.9033*
7	***0.0000***	***1.0000***	**0.2500**	**0.7998**	0.2500	0.7266	**0.0625**	**0.9952**	**0.0625**	**0.9923**	0.1875	0.7838	0.1250	**0.9902**
8	*0.0625*	0.8953	0.4375	0.7174	**0.1250**	**0.8478**	**0.0625**	***0.9952***	0.1875	0.7238	***0.0625***	**0.9747**	***0.0625***	0.9172
9	0.1875	0.7649	**0.2500**	0.7304	0.2500	0.7266	0.1875	0.7863	0.2500	0.6859	0.3125	0.5928	*0.1250*	*0.9172*
10	0.1875	0.7146	**0.2500**	0.7304	0.3125	0.6929	0.1875	0.7503	0.1875	0.7764	*0.1250*	*0.8783*	0.1875	0.8128

Table 2: Approximation of French full-book complexity using reduced samples. S denotes the number of samples. Performance is measured using Mean Absolute Error (MAE ↓) and Spearman's rank correlation coefficient (ρ ↑). Best values in each column are marked in bold. Best values in each row are marked in italic.

	Mean		Max		Min		Median		Mode		Q25		Q75	
S	MAE ↓	ρ ↑	MAE ↓	ρ ↑	MAE ↓	ρ ↑	MAE ↓	ρ ↑	MAE ↓	ρ ↑	MAE ↓	ρ ↑	MAE ↓	ρ ↑
1	0.2787	*0.8396*	0.4828	0.5847	0.7931	0.3970	*0.1379*	0.8333	*0.1379*	0.8333	0.3276	0.7352	0.2414	0.7701
2	0.2378	*0.8519*	0.3103	0.6708	0.7241	0.4181	0.1897	0.8260	0.2069	0.7723	0.4052	0.7033	*0.1466*	0.8425
3	0.1521	*0.9146*	0.3103	0.5787	0.5172	0.2940	*0.1379*	0.8487	*0.1379*	0.8487	0.2241	0.8138	0.1897	0.8368
4	0.1907	0.8978	0.2414	0.6920	0.4828	0.5670	0.1897	0.8334	0.2069	0.8044	0.2672	0.7810	*0.1034*	*0.9104*
5	0.1671	*0.9172*	0.2414	0.7335	0.2759	0.6002	*0.1379*	0.8261	*0.1379*	0.8261	0.2241	0.7246	0.2414	0.7824
6	*0.1109*	*0.9428*	0.2069	0.7273	0.2759	0.5287	0.1379	0.8678	0.1379	0.8341	0.1810	0.8619	0.1293	0.8488
7	0.0825	*0.9718*	0.1379	0.8051	0.3103	0.4862	***0.0345***	**0.9595**	***0.0345***	**0.9595**	0.1379	**0.9396**	**0.0517**	**0.9459**
8	0.1011	*0.9645*	**0.1034**	**0.8482**	0.2069	0.6772	0.1379	0.8816	0.1724	0.8218	0.1897	0.8923	*0.0603*	0.9271
9	*0.0912*	*0.9630*	**0.1034**	**0.8482**	0.2414	0.5666	0.1034	0.8880	0.1034	0.8880	0.1379	0.8790	0.1034	0.8974
10	**0.0699**	***0.9856***	0.1379	0.8051	**0.1034**	**0.7614**	0.0517	0.9571	***0.0345***	**0.9595**	**0.1034**	0.8852	0.0603	0.9057

Overall, $S = 7$ represents a practical trade-off between computational efficiency and approximation quality, as error is low, rank consistency is high, and systematic deviations are substantially reduced for robust aggregation methods, with only marginal gains observed beyond this point.

3.3. Experiment 2: Comparison with Human Gold Standard

The second experiment evaluates the correspondence between automatically estimated book-level complexity and a human-annotated gold standard. In this setting, expert annotators assigned complexity levels to sampled excerpts using the iR4S L1–L4 scale (Amaro et al., 2025; Monteiro et al., 2023), following established annotation guidelines.

The use of excerpt-based annotation reflects both practical constraints and real-world evaluation scenarios, where full-book assessment is typically performed through representative sampling.

Segment-level annotations were adjudicated to obtain a reliable reference, and inter-annotator agreement was measured to assess annotation consistency. Book-level complexity estimates were then derived by aggregating the annotated excerpt-level labels using the same set of aggregation strategies explored in Experiment 1. The strategy of estimating long-text complexity by aggregating excerpt–level complexity not only mimics the intended real-world scenario — where trainers/librarians assess a book's difficulty by reviewing a limited number of pages or passages–but also replicates the experimental setting of Experiment 1.

For Portuguese, inter-annotator agreement was measured on a subset of 37 excerpts (out of 150 total) that were independently labelled by two annotators. Because the complexity levels are ordinal (i.e., the ordering between classes is meaningful), we employed Cohen's weighted kappa coefficient, which accounts not only for agreement by chance but also for the degree of disagreement between ordered categories. The resulting kappa score of 0.733 indicates substantial agreement between an-

notators, according to commonly used interpretation scales. This level of agreement suggests that the annotation task is reasonably well-defined and consistently interpretable.

Performance is evaluated by comparing the automatically derived book-level estimates against the reference book-level labels inferred from human annotations. Metrics include accuracy, MAE, and agreement-oriented measures appropriate for ordinal scales. This experiment allows us to identify aggregation strategies that not only perform well in automatic settings but also align most closely with expert human judgment.

Although complexity levels form an ordinal scale, we report standard classification metrics (Accuracy, Precision, Recall, F1-score) to facilitate comparability with prior work in readability assessment. We complement these with error-based metrics (MAE) and agreement-oriented measures, which better capture the magnitude of deviations between predicted and reference levels. We acknowledge that future work could explore ordinal-specific evaluation measures.

The results of this comparison should provide critical validation for the proposed methodology and inform practical recommendations regarding sample size and aggregation functions for book-level complexity estimation. Again, the main objective of this experiment was to determine the most effective sampling strategy for estimating the global complexity level of a book. Since book-level complexity is inferred from annotations at the excerpt level, two key design choices were explored: (i) the size of the pool of excerpts sampled from each book, and (ii) the aggregation method used to combine the complexity annotations of these excerpts into a single book-level estimate.

To this end, we experimented with three different sample sizes (3, 5, and 10 evenly-distributed excerpts per book) and the same diverse set of aggregation strategies as before, including statistical operators (e.g., average, minimum, maximum, median), frequency-based aggregation, and quantile-based approaches. This setup allows us to assess how sensitive the final complexity estimation is to both the amount of available evidence and the way in which this evidence is summarised. Each configuration was evaluated using standard classification metrics (macro-averaged Accuracy, Precision, Recall, and F1-score), comparing the predicted global complexity level with the reference labels.

For the Portuguese books, a first observation from the results in Table 3 is that aggregation strategy has a stronger impact on performance than sample size. While increasing the number of excerpts from 3 to 10 sometimes leads to modest improvements, these gains are neither consistent nor systematic across aggregation methods. This suggests that simply sampling more excerpts does not necessarily yield a better estimate of global book complexity, and that the way excerpt-level information is combined is crucial.

Across all configurations, Accuracy values range between 0.25 and 0.50, indicating that the task remains challenging and sensitive to methodological choices. Precision, on the other hand, shows larger variability, particularly for aggregation methods that emphasise extreme values.

Aggregation methods based on minimum values (Min and Q25) consistently achieve the best overall performance. These methods reach the highest Accuracy (0.500) and Recall (0.500), and also obtain the best Precision scores (up to 0.711). This trend suggests that the least complex excerpts within a book are particularly informative for estimating its global complexity level. In other words, even a small number of simpler passages may significantly influence how the overall complexity of a book is perceived or categorised.

The Median aggregation strategy also performs competitively, especially for 5 samples, where it achieves an Accuracy of 0.500 and a strong F1-score of 0.490. This indicates that robust, central-tendency measures can provide a good balance between ignoring outliers and retaining representative information from the excerpt pool.

In contrast, Average aggregation yields moderate but stable results across different sample sizes. While it does not achieve the best scores on any metric, its performance is relatively insensitive to the number of excerpts considered. This stability may make it attractive in scenarios where robustness and simplicity are prioritised over peak performance.

Aggregation strategies based on maximum values consistently underperform. The Max strategy exhibits the lowest Accuracy, Precision, and F1-scores across almost all sample sizes. This suggests that focusing on the most complex excerpts leads to noisy or overly pessimistic estimates of global book complexity, likely because highly complex passages are not representative of the book as a whole.

The Mode strategy shows intermediate performance, with reasonable Precision but lower Recall and Accuracy, particularly as the sample size increases. As observed in Experiment 1, this may indicate that mode-based aggregation is sensitive to sampling variability and may struggle when the excerpt-level annotations are heterogeneous.

Finally, in this experiment, Q75 performs worse than its lower-quantile counterpart, reinforcing the suggestion that emphasising higher complexity excerpts does not yield reliable global estimates.

For the French books, an annotation task was carried out using the same guidelines as for the

Table 3: Results for different aggregation strategies and sample sizes of Portuguese books.

Aggregator	S	Accuracy	Precision	Recall	F1-score
Average	3	0.375	0.310	0.375	0.323
	5	0.438	0.357	0.438	0.379
	10	0.438	0.360	0.438	0.379
Max	3	0.250	0.161	0.250	0.195
	5	0.250	0.156	0.250	0.192
	10	0.312	0.179	0.312	0.227
Min	3	**0.500**	**0.711**	**0.500**	0.465
	5	0.438	0.558	0.438	0.398
	10	**0.500**	0.635	**0.500**	**0.493**
Median	3	0.438	0.573	0.438	0.430
	5	**0.500**	0.624	**0.500**	0.490
	10	0.438	0.569	0.438	0.419
Mode	3	0.375	0.538	0.375	0.389
	5	0.438	0.583	0.438	0.443
	10	0.312	0.487	0.312	0.322
Q25	3	**0.500**	**0.711**	**0.500**	0.465
	5	0.375	0.519	0.375	0.364
	10	**0.500**	**0.711**	**0.500**	0.465
Q75	3	0.438	0.573	0.438	0.430
	5	0.375	0.456	0.375	0.344
	10	0.375	0.485	0.375	0.343

Table 4: Results for different aggregation strategies and sample sizes of French books.

Agg.	S	Accuracy	Precision	Recall	F1-score
Mean	3	0.724	0.663	0.662	0.628
	5	0.655	0.523	0.454	0.484
	10	0.759	0.678	0.680	0.651
Max	3	0.793	**0.807**	0.728	0.755
	5	0.724	0.758	0.789	0.759
	10	0.621	0.711	0.644	0.643
Min	3	0.690	0.529	0.462	0.484
	5	0.793	0.549	0.527	0.536
	10	0.759	0.794	0.752	0.748
Median	3	0.724	0.663	0.662	0.628
	5	0.655	0.523	0.454	0.484
	10	0.759	0.678	0.680	0.651
Mode	3	0.759	0.683	0.833	0.651
	5	0.690	0.541	0.458	0.495
	10	0.793	0.696	0.850	0.678
Q25	3	0.586	0.497	0.399	0.442
	5	0.793	0.592	0.533	0.561
	10	0.793	0.575	0.525	0.547
Q75	3	**0.828**	0.781	**0.874**	**0.806**
	5	0.621	0.567	0.599	0.572
	10	0.724	0.763	0.708	0.652

Portuguese team. Inter-rater agreement was measured on a subset of 40 excerpts (out of 290 total). The resulting quadratic kappa score of 0.606 indicates a fair degree of agreement between the annotators, which then conducted a full annotation of the remaining excerpts. Finally, the annotators discussed a Gold Standard which would be used as a baseline to assess the model's performance.

Table 4 compares different aggregation strategies for combining excerpt-level predictions into book-level readability labels across varying numbers of sampled excerpts ($S \in \{3, 5, 10\}$). Overall, Q75 with $S = 3$ achieves the best performance in terms of Accuracy (0.828), Recall (0.874), and F1-score (0.806), while Max with $S = 3$ yields the highest Precision (0.807). These results suggest that upper-quantile–based aggregation strategies are particularly effective at capturing overall book-level difficulty. Interestingly, increasing the number of excerpts does not consistently improve performance. In several cases (e.g., Q75 and Max), performance decreases when moving from $S = 3$ to larger sample sizes. This indicates that a small number of informative excerpts may be sufficient to characterise the overall readability of a book, and that aggregating too many segments may introduce noise rather than additional signal. In contrast, central tendency measures such as Average and Median exhibit stable but consistently lower performance compared to upper-bound–oriented strategies (Max, Q75). This pattern suggests that book-level readability may be driven more strongly by the most difficult segments rather than by average difficulty. Overall, these findings indicate that readability at the book level behaves as a *peak-driven phenomenon*, where the most complex portions disproportionately influence global difficulty judgments.

3.4. Discussion

Regarding sampling-based automatic estimation, the results for both languages show that moderate sampling levels ($S = 6$-8) are sufficient to achieve low error, stable ranking, and no statistically significant deviation from the full-book baseline. In contrast, extreme-value methods remain comparatively unstable and systematically biased. Overall, $S = 7$ represents a practical trade-off between computational efficiency and approximation quality: error rates are low, rank consistency is high, and systematic deviations are substantially reduced for robust aggregation methods, with only marginal improvements observed beyond this point. Taken together, these findings demonstrate that central-tendency aggregation is highly robust under reduced sampling conditions.

The comparison with the human gold standard provides further insights for practical recommendations regarding both sample size and aggregation functions. In both languages, central-tendency estimators yield moderate yet stable results across different sample sizes, making them attractive for the practical scenarios considered. However, extreme-value estimators achieve the best performance in specific cases: in French, the Q75 with three samples yields the lowest MAE and the highest Spearman correlation; in Portuguese, Min and the Q25 suggest that book-level complexity may be shaped

by its most extreme sections rather than by passages that reflect the general level of the text. Nevertheless, given the level of inter-annotator agreement observed, these results may also reflect systematic tendencies among annotators to over- or underestimate the complexity level of the samples.

The two experiments provide complementary perspectives. Experiment 1 evaluates approximation to an automatic full-book profile and favours central-tendency aggregation with moderate sampling ($\approx$ 7 excerpts). In contrast, Experiment 2 evaluates alignment with human judgments, where quantile- and extreme-based strategies may capture perceptual difficulty more effectively. The final recommendation should, therefore, be understood as a practical compromise between stability, interpretability, and alignment with human perception.

Importantly, the proposed sampling framework is not intended as a substitute for full-text processing when such analysis is feasible, but rather as a robust approximation strategy under realistic constraints. These include limited access to full texts, computational cost considerations, and the need for interpretable and reproducible procedures in educational and library settings.

3.5. Book-Complexity Profiling

In light of the results discussed above, and consistent with the empirical finding that a limited number of well-distributed excerpts suffices to approximate global book complexity, we propose a lightweight, user-oriented sampling protocol for the real-world deployment of a database of book titles with automatically assigned complexity levels.[3]

Users are first asked to indicate the total number of pages of the text to be classified. If the text exceeds seven pages, the system requests seven evenly distributed samples; if it contains seven pages or fewer (e.g., a short article), users are asked to provide one sample per page. In all cases, each sample should contain approximately 250–300 words of running text, excluding paratextual material (e.g., table of contents, headers, footers, notes, images, captions, or references). Users are instructed to avoid truncating paragraphs and to indicate the page number of each excerpt. Samples should be drawn from the initial, middle, and final sections of the text to ensure even distribution and representativeness. Each excerpt is analysed automatically and assigned a predicted complexity level, which is stored together with its relative position in the text.

Once all required samples have been processed, the system computes the overall book-level complexity by averaging the excerpt-level predictions. It then presents both the individual sample classifications (the "book profile") and the aggregated result. Users may accept the proposed classification—thereby creating or updating the corresponding database entry, repeat the procedure with alternative samples, or terminate the process. The protocol is intentionally simple and resource-efficient, reflecting the empirical evidence that reliable book-level estimates can be obtained without exhaustive sampling.

[3] https://db.iread4skills.com/

4. Conclusion

This paper investigated principled strategies for estimating the global complexity level of full-length books in Portuguese and French under realistic sampling constraints. Departing from purely passage-oriented readability assessment, we examined how different sampling densities and aggregation operators affect the robustness of book-level complexity estimation.

Across both languages, results show that reliable global estimates can be obtained from a limited number of evenly distributed excerpts. Increasing the number of samples beyond a moderate threshold yields diminishing returns, indicating that lightweight sampling protocols are both feasible and methodologically sound. However, aggregation strategy plays a decisive role. While central-tendency measures (mean, median) provide stable approximations for automatic settings, aggregation functions that emphasise distributional extremes or upper quantiles prove particularly informative when aligning with human judgments, especially in French.

The findings support a distributional view of book-level readability, whereby global complexity emerges from the interplay between representative and peak difficulty segments. Importantly, the experiments demonstrate that computationally efficient, sampling-aware procedures can approximate full-text analysis without requiring complete textual access, making them suitable for deployment in Adult Learning scenarios and large-scale database construction.

A limitation of the present study lies in the relatively small number of books per language, which constrains the statistical generalisation of the findings and calls for further validation on larger and more diverse corpora.

Future work should further investigate cross-linguistic generalisation, annotation biases, and the interaction between sampling design and model architecture, as well as explore alternative sampling strategies, including randomised selection across the document, and the interaction between sampling design and book length. These aspects were beyond the scope of the present study but consti-

tute important directions for extending the proposed framework. Nonetheless, the present study provides empirical evidence and practical guidelines for moving beyond passages towards scalable and interpretable book-level readability assessment.

5. Acknowledgements

This work was supported by the European Commission (Project: iRead4Skills, Grant number: 1010094837, Topic: HORIZON-CL2-2022-TRANSFORMATIONS-01-07, DOI: 10.3030/101094837), and by Portuguese national funds through FCT (References: UID/50021/2025 (DOI: 10.54499/UID/50021/2025), UID/PRR/50021/2025 (DOI: 10.54499/UID/PRR/50021/2025) and UID/03213/2025 - CLUNL, DOI: 10.54499/UID/03213/2025). We gratefully acknowledge the contribution of the human annotators of the Portuguese corpus: S. Barbosa, R. Monteiro, I. Müller, M. Moutinho, and S. Reis.

5.1. Ethical considerations

The iRead4Skills system[4] ensures a proportionate, legally sound, and socially beneficial use of technology in support of literacy and the European publishing ecosystem. It does not infringe copyright or prejudice the normal exploitation of works.

The system performs automated analysis of texts solely for the purpose of assessing text complexity, involving at most temporary and technically necessary acts of reproduction. The submission of titles to the iR4S database is carried out through a back-office interface accessible only to authorised publishers, librarians, and project team members. Submitted text excerpts are never displayed or made publicly accessible and are immediately discarded after processing. Extracted text is processed in a stateless computational environment, held only transiently in working memory for textual and linguistic analysis, and automatically deleted upon completion of the process. No text, excerpts, images, or other expressive elements of the work are stored, indexed, cataloged, or transmitted. The only persistent output is a non-expressive metadata label indicating the degree of complexity associated with the excerpts and the corresponding title.

Under the law[5], acts of reproduction that are transient or incidental, that constitute an integral and essential part of a technological process, and that have no independent economic significance fall outside the scope of the reproduction right.

Non-textual information—such as bibliographic metadata, the assigned complexity level, and the identifier of the contributing user—may be retained for the purposes of tracking, auditing, and analysis.

Users may, at any time, elect not to include a given title and its associated complexity assessment in the database or request their deletion.

6. Lay summary

Choosing books for adults with low reading skills is not easy. This is especially true when we want to judge the difficulty of a whole book, not just a short text. A few short passages do not always show how easy or hard the full book really is.

In this paper, we study how to estimate the difficulty of full books in Portuguese and French. We test different ways of choosing short passages from a book and combining the results. We also use computer tools to analyse language and estimate how hard a text is to read.

The books are grouped into four difficulty levels, from very easy (L1) to more complex (L4). This scale was designed for adult learners and follows widely used international standards.

Our goal is to help build a public database of books with reliable difficulty information. This can help adult learners, teachers, librarians, and publishers choose reading materials that better match readers' needs.

7. Bibliographical References

Wafa Aissa, Raquel Amaro, David Antunes, Thibault Bañeras-Roux, Jorge Baptista, Alejandro Catala, Luís Correia, Thomas François, Marcos Garcia, Mario Izquierdo-Álvarez, Nuno Mamede, Vasco Martins, Miguel Neves, Eugénio Ribeiro, Sandra Rodriguez Rey, and Elodie Vanzeveren. 2025a. The iRead4Skills intelligent complexity analyzer. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 73–84, Suzhou, China. Association for Computational Linguistics.

Wafa Aissa, Thibault Bañeras-Roux, Elodie Vanzeveren, Lingyun Gao, Rodrigo Wilkens, and Thomas François. 2025b. Assessing French readability for adults with low literacy: A global and local perspective. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20517–20539, Suzhou, China. Association for Computational Linguistics.

[4] https://v1.iread4skills.com/

[5] Directive 2001/29/EC: https://eur-lex.europa.eu/eli/dir/2001/29/oj/eng)

Simon Allan, Marie McGhee, and Rob van Krieken. 2005. Using readability formulae for examination questions. Technical report, Qualifications and Curriculum Authority.

Raquel Amaro, Susana Correia, Ricardo Monteiro, Alice Pintard, Michell Moutinho, and Sílvia Barbosa. 2025. Framework of textual complexity for low-literacy adults: Levels and descriptors within the iread4skill project. *Langues & Parole*, 10:57–119.

I. Beltagy, M. E. Peters, and A. Cohan. 2020. Longformer: The long-document transformer. *arXiv: Computation and Language*, pages 1–17.

Rebekah G. Benjamin. 2012. Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, 24(1):63–88.

Jeanne S. Chall and Edgar Dale. 1995. Readability revisited: The new dale–chall readability formula. *Brookline Books*.

K. Collins-Thompson and J. Callan. 2005. Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science and Technology*, 56(13):1448–1462.

Kevyn Collins-Thompson. 2014. Computational assessment of text readability: A survey of current and future research. *ITL – International Journal of Applied Linguistics*, 165(2):97–135.

Council of Europe. 2020. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment – Companion Volume*. Council of Europe Publishing, Strasbourg.

Scott A. Crossley, Andrea Heintz, Jae Sung Choi, Jesse Batchelor, Mehdi Karimi, and Adam Malatinszky. 2023. A large-scaled corpus for assessing text readability. *Behavior Research Methods*, 55:491–507.

Edgar Dale and Jeanne S. Chall. 1948. A formula for predicting readability. *Educational Research Bulletin*, 27(1):11–20, 28.

Tovly Deutsch, Masoud Jasbi, and Stuart M Shieber. 2020. Linguistic features for readability assessment. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–17.

Lijun Feng, Noémie Elhadad, and Matt Huenerfauth. 2010. Comparison of features for automatic readability assessment. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, pages 276–284.

Rudolf Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

T. François and C. Fairon. 2012. An "AI readability" formula for French as a foreign language. In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing (EMNLP 2012)*, pages 466–477.

B. W. Lee, Y. S. Jang, and J. Lee. 2021. Pushing on text readability assessment: A transformer meets handcrafted linguistic features. In *EMNLP 2021*, pages 10669–10686.

Wenbiao Li, Rui Sun, Tianyi Zhang, and Yunfang Wu. 2024. Going beyond passages: Readability assessment for book-level long texts. In *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 1: Main Conference)*, pages 1298–1309, Taiyuan, China. Chinese Information Processing Society of China.

G. Harry McLaughlin. 1969. Smog grading—a new readability formula. *Journal of Reading*, 12(8):639–646.

OECD. 2013a. *The Survey of Adult Skills: Reader's Companion*. OECD Publishing, Paris, France.

OECD. 2013b. Technical report of the survey of adult skills (piaac). Technical report, OECD Publishing, Paris, France.

OECD. 2021. The assessment frameworks for cycle 2 of the programme for the international assessment of adult competencies. Technical report, OECD Publishing, Paris.

Sarah E. Petersen and Mari Ostendorf. 2009. A machine learning approach to reading level assessment. *Computer Speech & Language*, 23(1):89–106.

Eugénio Ribeiro, Nuno Mamede, and Jorge Baptista. 2024a. Avaliação automática do nível de complexidade de textos em português europeu. *Linguamática*, 16(2):115–139.

Eugénio Ribeiro, Nuno Mamede, and Jorge Baptista. 2024b. Text readability assessment in european portuguese: A comparison of classification and regression approaches. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese (PROPOR 2024)*, pages 551–557.

Eugénio Ribeiro, David Antunes, Nuno Mamede, and Jorge Baptista. 2025. Exploring Few-Shot Approaches to Automatic Text Complexity Assessment in European Portuguese. *Journal of the Brazilian Computer Society*, 31(1):690–710.

Sarah Schwarm and Mari Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL '05)*, pages 523–530, Ann Arbor, Michigan. Association for Computational Linguistics.

K. M. Sheehan, I. Kostin, Y. Futagi, and M. Flor. 2010. Generating automated text complexity classifications that are aligned with targeted text complexity standards. *ETS Research Report Series*, 2:1–44.

A. Jackson Stenner, Hal Burdick, Elizabeth E. Sanford, and David S. Burdick. 2006. How accurate are lexile text measures? *Journal of Applied Measurement*, 7(3):307–322.

Rodrigo Wilkens, Patrick Watrin, Rémi Cardon, Alice Pintard, Isabelle Gribomont, and Thomas François. 2024. Exploring hybrid approaches to readability: experiments on the complementarity between linguistic features and transformers. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2316–2331.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems*, volume 33, pages 17283–17297. Curran Associates, Inc.

8. Language Resource References

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Ricardo Monteiro, Raquel Amaro, Susana Correia, Alice Pintard, Roser Gauchola, Michell Moutinho, and Xavier Blanco Escoda. 2023. iRead4Skills - Complexity Levels. Technical report, Zenodo. Version 1.0.

Alice Pintard, Thomas François, Justine Nagant de Deuxchaisnes, Sílvia Barbosa, Maria Leonor Reis, Michell Moutinho, Ricardo Monteiro, Raquel Amaro, Susana Correia, Sandra Rodríguez Rey, Marcos Garcia González, Keran Mu, and Xavier Blanco Escoda. 2024. iRead4Skills Dataset 1: Corpora by Complexity Level for FR, PT and SP. Technical report, Zenodo. Version 2.1.

João Rodrigues, Luís Gomes, João Silva, António Branco, Rodrigo Santos, Henrique Lopes Cardoso, and Tomás Osório. 2023. Advancing Neural Encoding of Portuguese with Transformer Albertina PT-*. In *EPIA Conference on Artificial Intelligence*, pages 441–453. Springer.

A. Portuguese corpus

This appendix provides an overview of the Portuguese corpus (Table 5), reports the excerpt-level and aggregated complexity annotations for all extracted samples (Table 6), and presents the relationship between sampling density and approximation accuracy in terms of MAE and Spearman's ρ (Figure 1).

B. French corpus

This appendix provides an overview of the French corpus (Table 7), reports the excerpt-level and aggregated complexity annotations for all extracted samples (Table B), and presents the relationship between sampling density and approximation accuracy in terms of MAE and Spearman's ρ (Figure 2).

C. Technical details

The complexity assessment model used for Portuguese (Ribeiro et al., 2025) is a fine-tuned version of the smallest Albertina PT-PT model (Rodrigues et al., 2023), a transformer-based European Portuguese encoder. The model's classification strategy bases predictions on the weighted average of the probability distribution in contrast to simply selecting the most probable class.

For French passage readability assessment, we use the model proposed by Aissa et al. (2025a,b). The model is based on CamemBERT (Martin et al., 2020) and fine-tuned on a French corpus specifically designed for adults with low literacy levels.

Table 5: Portuguese corpus. Fragments from the texts with '*' have also been included in the iRead4Skills corpus (Pintard et al., 2024).

ID	Title	Level	#word	%
Text-01	A minha cidade é um livro	1	619	0,07
Text-02	Miguel e Sinatra	1*	3 205	0,34
Text-03	O paraiso sao os outros	1*	2 201	0,23
Text-04	O triunfo dos porcos	1	35 005	3,71
Text-05	Pageboy	2*	75 832	8,04
Text-06	O Principezinho	2	9 053	0,96
Text-07	Programa BE	2*	7 011	0,74
Text-08	Todos devemos ser feministas	2	11 129	1,18
Text-09	1984	3*	96 091	10,19
Text-10	Conto de fadas	3*	217 330	23,04
Text-11	Na sombra principe Harry	3*	162 634	17,24
Text-12	Programa ADN	3*	56 312	5,97
Text-13	Almoco de domingo	4*	58 697	6,22
Text-14	Mil anos de alegrias e tristezas	4*	120 766	12,80
Text-15	O remorso de Baltazar Serapião	4*	52 086	5,52
Text-16	Salario médio em Portugal	4*	35 198	3,73
		total	943 169	

Table 6: Excerpt-level and aggregated annotated complexity levels for each Portuguese text considering all samples extracted. M1 stands for Average, M2 stands for Max, M3 stands for Min, M4 stands for Median, M5 stands for Mode, M6 stands for Q25, and M7 stands for Q75.

	Excerpts										Overall Complexity (by metric)						
Book	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	M1	M2	M3	M4	M5	M6	M7
Text-01	1	2	X	X	X	X	X	X	X	X	2	2	1	1	1	1	1
Text-02	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Text-03	1	3	3	2	2	2	2	2	X	X	2	3	1	2	2	3	3
Text-04	3	3	2	3	3	3	3	2	3	3	3	3	2	3	3	2	3
Text-05	3	2	3	3	2	3	3	3	2	2	3	3	2	3	3	2	3
Text-06	2	2	2	2	3	3	2	3	3	2	2	3	2	2	2	2	3
Text-07	1	2	3	3	3	3	2	2	3	2	2	3	1	2	3	2	3
Text-08	3	3	2	3	3	3	2	2	2	3	3	3	2	3	3	3	3
Text-09	3	3	3	3	3	4	3	4	3	3	3	4	3	3	3	3	3
Text-10	2	2	2	2	3	3	3	3	4	4	3	4	2	3	2	2	3
Text-11	2	3	3	3	3	2	3	2	3	3	3	3	2	3	3	2	3
Text-12	3	4	4	4	3	4	3	4	3	4	4	4	3	4	4	3	4
Text-13	3	2	4	3	3	2	2	3	2	2	3	4	2	2	2	2	3
Text-14	3	4	4	4	4	3	4	3	3	3	4	4	3	3	3	3	4
Text-15	3	4	3	3	3	4	3	4	3	3	3	4	3	3	3	3	3
Text-16	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4

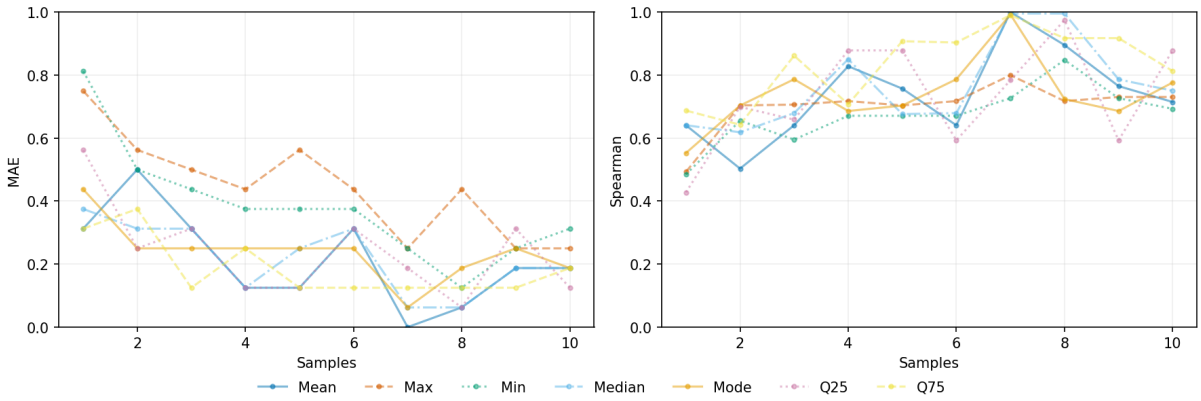


Figure 1: MAE and Spearman's ρ as a function of the number of sampled excerpts (S) for Portuguese books.

Table 7: French corpus.

ID	Title	Publisher	#word	%
Text-01	15 jours pour réussir	Didier	7 701	3.18%
Text-02	5 Contes	Hachette	15 464	6.39%
Text-03	Carmen	Hachette	11 803	4.88%
Text-04	Carmen	De Boeck	7 577	3.13%
Text-05	Contes	Hachette	10 405	4.30%
Text-06	Lettres de mon moulin	De Boeck	4 783	1.98%
Text-07	Fantôme de l'Opéra	De Boeck	10 067	4.16%
Text-08	Germinal	Hachette	12 727	5.26%
Text-09	Julie est amoureuse	Hachette	3 696	1.53%
Text-10	La boîte en os	De Boeck	3 524	1.46%
Text-11	La disparition	Hachette	4 008	1.66%
Text-12	La fille qui vivait hors du temps	Didier	8 694	3.59%
Text-13	La nuit blanche de Zoé	Hachette	3 828	1.58%
Text-14	La tête d'un homme	Hachette	17 421	7.20%
Text-15	Lancelot	De Boeck	5 817	2.40%
Text-16	Le Roi Arthur	De Boeck	3 241	1.34%
Text-17	Le blog de Maia	Hachette	3 868	1.60%
Text-18	Le casque mystérieux	Didier	7 414	3.06%
Text-19	Le prisonnier du temps	Hachette	3 179	1.31%
Text-20	Le secret du vieil orme	De Boeck	8 159	3.37%
Text-21	Les Trois Mousquetaires	De Boeck	14 322	5.92%
Text-22	Les Misérables	Hachette	18 172	7.51%
Text-23	Lucas sur la route	Hachette	3 280	1.36%
Text-24	Maigret tend un piège	Hachette	18 595	7.68%
Text-25	Le mystère de la chambre jaune	De Boeck	12 213	5.05%
Text-26	Double assassinat dans la Rue Morgue	De Boeck	7 336	3.03%
Text-27	La Lettre volée	De Boeck	3 051	1.26%
Text-28	Tristan et Iseult	De Boeck	6 784	2.80%
Text-29	Une étrange disparition	De Boeck	4 885	2.02%
	Total		242 014	100.00%

Table 8: Excerpt-level and aggregated annotated complexity levels for each French books considering all samples extracted. M1 stands for Average, M2 stands for Max, M3 stands for Min, M4 stands for Median, M5 stands for Mode, M6 stands for Q25, and M7 stands for Q75.

	Excerpts										Overall Complexity (by metric)						
Book	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	M1	M2	M3	M4	M5	M6	M7
Text-01	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Text-02	3	3	2	2	3	3	3	3	3	3	3	3	2	3	3	3	3
Text-03	3	3	3	3	3	3	2	2	2	3	3	3	2	3	3	2	3
Text-04	4	3	3	3	3	4	4	3	4	4	4	4	3	4	3	3	4
Text-05	2	3	2	2	2	2	2	2	2	2	2	3	2	2	2	2	2
Text-06	3	2	3	3	3	3	3	3	3	3	3	3	2	3	3	3	3
Text-07	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
Text-08	3	2	3	3	3	3	3	3	3	3	3	3	2	3	3	3	3
Text-09	3	2	3	2	2	2	2	2	2	2	2	3	2	2	2	2	2
Text-10	3	3	3	3	3	3	3	3	4	3	3	4	3	3	3	3	3
Text-11	2	2	2	2	3	2	2	2	2	3	2	3	2	2	2	2	2
Text-12	2	3	2	2	2	3	3	3	3	3	3	3	2	3	3	2	3
Text-13	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Text-14	3	3	2	3	2	3	2	2	3	3	3	3	2	3	3	2	3
Text-15	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
Text-16	3	3	3	3	3	3	4	3	3	3	3	4	3	3	3	3	3
Text-17	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Text-18	2	2	2	2	2	2	3	3	3	2	2	3	2	2	2	2	3
Text-19	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
Text-20	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
Text-21	3	4	3	3	3	3	4	4	3	3	3	4	3	3	3	3	4
Text-22	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
Text-23	3	2	2	3	3	3	3	3	2	2	3	3	2	3	3	2	3
Text-24	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
Text-25	4	3	3	3	4	4	3	3	3	3	3	4	3	3	3	3	4
Text-26	4	4	4	4	3	4	4	3	3	4	4	4	3	4	4	3	4
Text-27	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
Text-28	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
Text-29	2	2	2	2	2	3	2	3	3	3	2	3	2	2	2	2	3

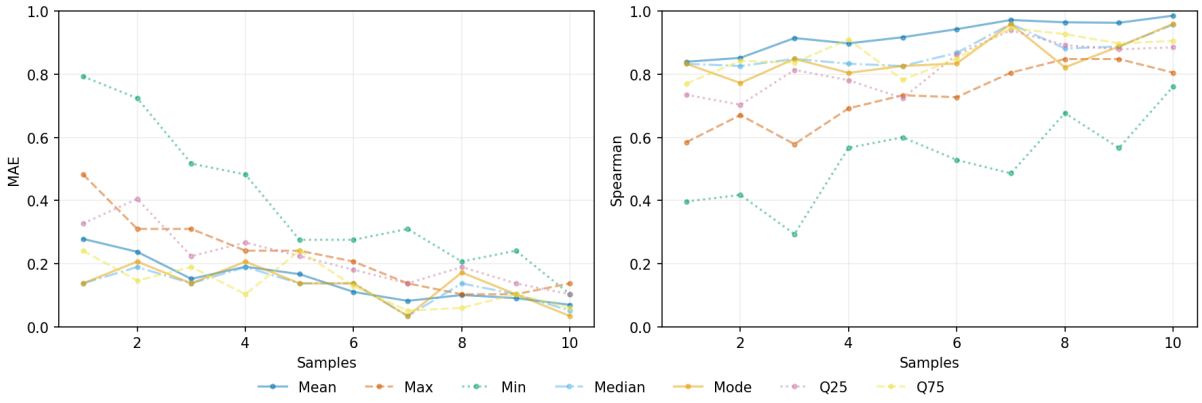


Figure 2: MAE and Spearman's ρ as a function of the number of sampled excerpts (S) for French books.

Taming CATS: Controllable Automatic Text Simplification through Instruction Fine-Tuning with Control Tokens

Hanna Hubarava[1,2] **Yingqiang Gao***[1]
[1]Department of Computational Linguistics, University of Zurich, Switzerland
[2]Department of Clinical Research, University of Bern, Switzerland
hanna.hubarava@unibe.ch, yingqiang.gao@cl.uzh.ch

Abstract

Controllable Automatic Text Simplification (CATS) produces user-tailored outputs, yet controllability is often treated as a decoding problem and evaluated with metrics that are not reflective to the measure of control. We observe that controllability in ATS is significantly constrained by *data and evaluation*. To this end, we introduce a domain-agnostic CATS framework based on instruction fine-tuning with discrete control tokens, steering open-source models to target readability levels and compression rates. Across three model families with different model sizes (Llama, Mistral, Qwen; 1–14B) and four domains (medicine, public administration, news, encyclopedic text), we find that smaller models (1–3B) can be competitive, but reliable controllability strongly depends on whether the training data encodes sufficient variation in the target attribute. Readability control (FKGL, ARI, Dale-Chall) is learned consistently, whereas compression control underperforms due to limited signal variability in the existing corpora. We further show that standard simplification and similarity metrics are insufficient for measuring control, motivating error-based measures for target-output alignment. Finally, our sampling and stratification experiments demonstrate that naive splits can introduce distributional mismatch that undermines both training and evaluation.

Keywords: Automatic Text Simplification, Controllable Text Generation, Instruction Fine-Tuning.

1. Introduction

Automatic Text Simplification (ATS) aims to reduce linguistic complexity while preserving meaning in order to enable accessibility of information for diverse purposes and audiences (Shardlow, 2014; Grabar and Saggion, 2022; Espinosa-Zaragoza et al., 2023). Recent advances in large language models (LLMs) have rekindled interest in *controllable* simplification (Kew et al., 2023; Tran et al., 2025), where systems are expected to adapt outputs to user-specified complexity levels rather than produce a static simplified variant. Controllability in ATS tends to be treated as a decoding problem (Martin et al., 2020), with a focus on conditioning mechanisms such as model prompting or inference configurations, while largely assuming that datasets, splits, and evaluation metrics reflect the intended notion of control. Yet, ATS datasets vary widely in how complexity is encoded, many exhibit minimal variation along key attributes (e.g., compression). Training LLMs on datasets of limited variation (Vásquez-Rodríguez et al., 2021) would thus fail to learn distinguish fine-grained controllability needs.

In this work, we present our approach to CATS: an instruction fine-tuning (IFT) framework using discrete control tokens with open-source decoder-only LLMs. We compared control effectiveness across five readability attributes (FKGL (Kincaid et al., 1975); ARI (Smith and Senter, 1967); Dale-Chall (Dale and Chall, 1948; Chall and Dale, 1995) and two compression levels (word and character), on four domain-specific datasets (MED-EASI (Basu et al., 2023); SIMPA (Scarton et al., 2018); WIKI-LARGE (Zhang and Lapata, 2017); NEWSELA (Xu et al., 2015)), by testing various model families (Llama (Dubey et al., 2024); Mistral (Jiang et al., 2023); Qwen (Yang et al., 2025)) and sizes (1-14B).

We argue that *data and evaluation* are equally important for CATS: our results show that LLMs can learn to target absolute readability levels through fine-tuning, but only when the training data contains sufficient and well-distributed learning signal. We also conducted extensive data experiments on sampling and stratification, and observed that use of native dataset splits or naively randomized partitioning can magnify distributional mismatch between training and evaluation sets. We further emphasize that metrics which take into account the error between target (reference simplification) and prediction (model output) values are indispensable for measuring controllability, since traditional simplification and similarity metrics are agnostic to target complexity deviations. With these findings, we show that for effective CATS solutions, robustness, reproducibility, and evaluation design are as critical as architectural choices.

2. Related Work

ATS is a specialized form of controllable text generation, traditionally modeled as monolingual machine translation (Wubben et al., 2012; Wang et al., 2016b; Sheang and Saggion, 2021). While it

*Corresponding author. Data Code

Proceedings of the Joint Workshop on Readability and Text Simplification (READIxTSAR) @ LREC 2026, pages 26–48
11 May 2026.

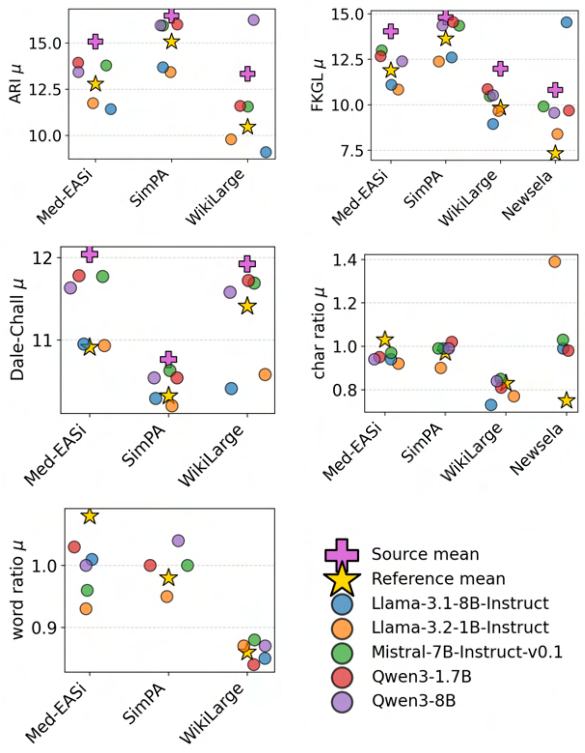


Figure 1: Mean control-attribute values across models and datasets. Fine-tuned models generate text that is simpler than the source but more complex than the target. Sentence-aligned datasets (MED-EASI, SIMPA, WIKILARGE) show almost no compression signal, limiting training and evaluation.

partly shares goals with text summarization (Alva-Manchego et al., 2020), ATS is distinct in its focus on readability and complexity reduction; simplifications may actually increase text length through explanatory paraphrasing or addition of cohesive markers (Alva-Manchego et al., 2020). Over the years, ATS research has transitioned from rule- and dictionary-based as well as statistical systems (Chandrasekar and Bangalore, 1997; Espinosa-Zaragoza et al., 2023) to data-driven neural approaches (Wang et al., 2016a; Nisioi et al., 2017; Zhang and Lapata, 2017), but the scarcity of high-quality parallel corpora remains a key bottleneck (Vásquez-Rodríguez et al., 2021; Agrawal and Carpuat, 2024).

Controllability in the ATS is guided by specific transformation attributes, which can be indicated in the instruction prompt or encoded in the control token. ACCESS framework (Martin et al., 2020) pioneered conditioning models on lexical complexity, length, and syntactic markers. Subsequent work has expanded this to proficiency levels (e.g., CEFR tokens (Spring et al., 2021)) and the mitigation of "copying behavior" through explicit labels (Sheang and Saggion, 2021). To bypass data scarcity, unsupervised approaches like MUSS (Martin et al., 2022) leverage mined paraphrases and unsupervised pre-training. Recently, controlled decoding methods such as FUDGE (Yang and Klein, 2021; Kew and Ebling, 2022) have been combined with paraphrase models to nudge outputs toward target complexity levels without requiring massive in-domain parallel data.

Faced with the inherent subjectivity of simplification quality evaluation (Grabar and Saggion, 2022), ATS evaluation typically considers adequacy (meaning preservation), fluency, and simplicity. Early work relied on machine translation metrics such as BLEU (Papineni et al., 2002), which are insensitive to structural transformations. SARI (Xu et al., 2016), now a standard metric for simplification, explicitly models lexical edit operations but has been shown to correlate weakly with human judgments in cases involving structural changes such as sentence splitting or merging (Alva-Manchego et al., 2020). Learned metrics such as BERTScore (Zhang et al., 2019) improve the assessment of semantic adequacy (Alva-Manchego et al., 2021), but remain general-purpose and do not explicitly account for simplicity. More recently, LENS (Maddela et al., 2023) has been introduced as a simplification-specific learned metric that holistically models human judgments of simplification quality, reflecting adequacy, fluency, and simplicity.

3. Methods

3.1. Control Attributes

We measured controllability in terms of model's ability to generate a text simplification of a specific control attribute value. We operated with five such numerical attributes: three readability attributes include FKGL (Kincaid et al., 1975), ARI (Smith and Senter, 1967), Dale-Chall (Dale and Chall, 1948; Chall and Dale, 1995), and two compression ratios in terms of character and word count.

3.2. Datasets

We curated a multi-domain corpus to evaluate the effectiveness of instruction fine-tuning for controllable text simplification in medicine, public administration, news, and encyclopedic knowledge, representing different text styles and linguistic characteristics. We used the following sources:

- MED-EASI (Basu et al., 2023). **Domain**: medical. **Level**: sentence. **Creation:** human annotations by medical experts and lay crowd-workers (Toloka), with AI-assisted workflow. **Mapping:** 1-to-1 complex→simple. **Simplification strategies:** token-level spans for *elaboration*, *replacement*, *deletion*, *insertion*.

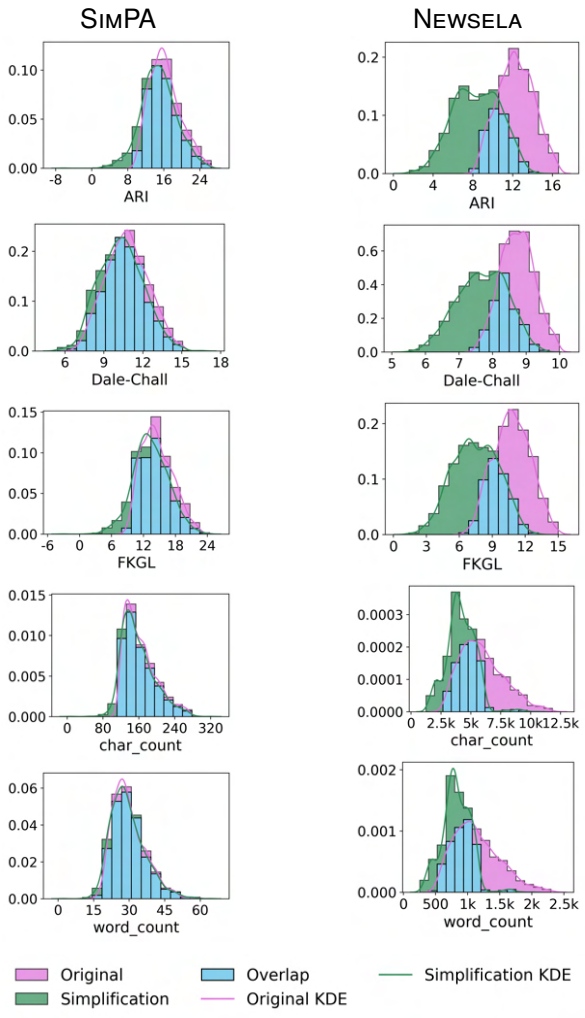


Figure 2: Both SimPA and Newsela show larger distribution shifts by readability than by length transformations. Text-aligned Newsela shows greater distribution shift than the sentence-aligned SimPA.

- SimPA (Scarton et al., 2018). **Domain**: public administration, **Level**: sentence. **Creation:** human simplification in two stages. **Mapping:** lexical subset is 1-to-3 (three lexical simplifications per complex source); syntactic subset is 1-to-1 (one lexically simplified version further simplified syntactically). **Simplification strategies:** explicit *lexical* vs. *syntactic* simplification.

- WikiLarge (Zhang and Lapata, 2017). **Domain**: general/encyclopedic, **Level**: sentence. **Creation:** automatic alignment between English Wikipedia and Simple English Wikipedia using similarity heuristics and filtering/cleaning. **Mapping:** predominantly 1-to-1; the native test set (not used by us) provides multiple references (8 per complex source). **Simplification strategies:** not explicitly annotated; transformations arise from mined revision/alignment pairs.

- Newsela (Xu et al., 2015). **Domain**: news. **Level**: document. **Creation:** professional editors rewrite articles for children at multiple grade levels. **Mapping:** 1-to-N (up to 4 simplifications per complex source, ordered by decreasing complexity). **Simplification strategies:** Edits involve content rewrites and frequent sentence splitting.

To better understand the learning signal available for different control attributes, we analyze their distributions across datasets (Fig. 2). We observe that the document-aligned Newsela dataset shows greater distributional spread across all control attributes than the sentence-aligned datasets.

3.3. Data Preprocessing Pipeline

Harmonization and Metric Calculation. All datasets were converted into a unified JSON Lines (JSONL) schema that enables consistent downstream processing. Each entry in the standardized format contains global metadata, source metrics and simplifications array. Global metadata includes: instance id, source text, dataset name, domain and language, annotation type, alignment level, and native split (if applicable). Source metrics include readability values, as well as character and word count. The simplifications array contains one or more simplifications each with the following information (if available, otherwise -1): simplification text, version, compression rates, target control-attribute values and similarity metric values.

The hierarchical JSONL format (one complex source text with multiple text simplifications) was flattened to create individual training instances. For datasets containing multiple reference simplifications for a single source, we converted these into individual complex-simple pairs.

Our approach is based on absolute control ("simplify to FKGL 5 level"), as opposed to relative control ("reduce complexity by X points"). Control attributes are automatically extracted from the dataset. For readability metrics, we computed source (complex) and reference (simplification) values rounded to the nearest integer (<FKGL=5>). For structural attributes, we computed reference/source length ratios in terms of the number of characters and words. The ratios were rounded to one decimal place (<WORD_COMPRESSION=0.5>).

Subsampling and Stratification. To ensure representative train/validation/test splits (80/10/10) across varying text complexities, we employed careful stratified sampling. We removed extreme outliers and filtered out texts falling below the 1st or above the 99th percentile for FKGL, ARI, Dale-Chall, and character length. To determine the optimal split, we utilized the Kolmogorov–Smirnov (KS) goodness-of-fit test. By selecting the partitioning

strategy that minimized the KS distance, we ensured that the validation and test sets are statistically representative of the training data. FKGL is our primary stratification feature selected as described in Section 5.1. We applied stratified sampling by FKGL to keep 3k complex-simple pairs for NEWSELA and 2k for WIKILARGE. Due to the mixed lexical-syntactic strategy employed in SIMPA, we merged these subsets by: (1) retaining all unique source (complex) sentences from each subset, and (2) for overlapping sentences, randomly assigning them to either the lexical or syntactic subset with 50-50 probability. This ensures no duplicate complex texts while preserving both simplification types.

Data filtering. To create the filtered subsets, we applied metric-based filtering that retains only instances where all readability values decrease from source (complex) to reference (simple). Instances where complexity increases or remains unchanged were excluded, as they do not reflect the monotonic readability assumption typically associated with simplification. This procedure creates monotonic (mono) variants of each split, with removal statistics logged per dataset and metric. Removal rates range from 10–30%, depending on dataset characteristics, with automatically aligned corpora (e.g., WIKILARGE) exhibiting higher removal rates than manually curated datasets.

4. Fine-tuning Pipeline

We fine-tuned the model such that it learns to generate a simplification whose absolute readability value (FKGL, ARI, Dale-Chall) or relative length ratio closely match a control value. Dataset simplifications served both as ground-truth references and sources of the target control attribute value.

4.1. Prompt Construction

To reduce over-reliance on a single phrasing, we utilized six manually curated system prompt variants, with randomly selection during fine-tuning. As shown in Fig. 3, the instruction prompt contains the complex source text and the reference control attribute. We employed a dynamic prompting strategy to integrate control attributes into the fine-tuning process. We used embedded control tokens in the format <METRIC=VALUE> prepended to the assistant's response. Values were precomputed based on reference simplifications and rounded to one decimal place. To maintain consistency with the models' pre-training, all prompts were formatted using model-native **chat template** via Hugging Face's `tokenizer.apply_chat_template()`. See Appendix E for technical details.

```
<|begin_of_text|>
<|start_header_id|>system<|end_header_id|>

You are a helpful expert in text simplification.
You generate a simplified version of the text input
by the user.  You simplify the text according to the
instructions given by the user.  When asked to
simplify a text, generate only the requested
simplification, without any additional comments,
notes or explanations.

<|eot_id|>
<|start_header_id|>user<|end_header_id|>

INSTRUCTION: Simplify the following text such that
its Flesch-Kincaid Grade Level (FKGL) score is
approximately equal to that specified in the control
token prepended to your generated simplification.
The control token has the following format:
<METRIC=VALUE>.

SOURCE TEXT: <FKGL=4.8> No cure for the common cold
exists, but the symptoms can be treated.

EXPLANATION: The <FKGL=4.0> token specifies that the
target Flesch-Kincaid Grade Level should be
approximately 4.0.  Lower values indicate simpler
text.

<|eot_id|>

<|start_header_id|>assistant
<|end_header_id|>
<FKGL=4.0><|eot_id|>
```

Figure 3: Blue: model-native automatic prompt formatting. Green: system prompt. Purple: source text with its control attribute value. Plum: target (reference) control attribute value.

4.2. Models

We conducted experiments across three open-source model families spanning the 1B–14B parameter range. The model lineup includes instruct models from three families: Llama (Dubey et al., 2024) (3.2-1B, 3.2-3B, 3.1-8B, 2-13B), Mistral (Jiang et al., 2023) (Ministral-3B, Mistral-7B-v0.3), and Qwen (Yang et al., 2025) (Qwen3-1.7B, -4B, -7B, -14B). For models exceeding 4B parameters, we used LoRA (Hu et al., 2021) to manage computational constraints.

4.3. Training Objective

By including the control token at the start of the assistant's turn, the model learns the conditional relationship between the specified metric value and the linguistic features of the generated simplification. We use the standard cross-entropy loss for training and validation, refraining from additional signals (e.g. based on a prediction-reference error) and keep training as a pure causal language modeling task. We mask the prompt tokens, and only the completion contributes to the loss.

4.4. Evaluation Metrics

We evaluate the models across the three distinct dimensions of controllability, simplification quality, and textual similarity. Controllability is measured

<FKGL=12.3> *Source* Under optimal conditions, it can destroy an entire orchard in a single growing season.
<FKGL=9.5> *Reference* If the conditions are right, it can destroy an entire orchard in a single growing season.
<FKGL=7.6> *Qwen3-8B* It can destroy an entire orchard in a single growing season.
<FKGL=8.4> *Ministral-3B-Instruct* In the perfect conditions, it can destroy an entire orchard in one season.
<FKGL=9.2> *Llama-3.2-1B-Instruct* The fungus can destroy an entire orchard in a single growing season.
<FKGL=9.2> *Llama-3.1-8B-Instruct* The fungus can destroy an entire orchard in a single growing season.
<FKGL=12.3> *Mistral-7B-Instruct-v0.1* Under optimal conditions, it can destroy an entire orchard in a single growing season.
<FKGL=12.3> *Qwen3-1.7B* Under optimal conditions, it can destroy an entire orchard in a single growing season.

Table 1: Dataset: MEDEASI. Control attribute: FKGL. Some models copy the source sentence, resulting in no simplification, while others successfully generate simplifications matching the target FKGL. Even minimal lexical changes can result in significant readability value shifts.

using Mean Absolute Error (MAE) between the target and prediction. Simplification quality is assessed via SARI (Xu et al., 2016) and LENS (Maddela et al., 2023), additionally reporting COMET (Rei et al., 2020) as a general-purpose semantic similarity metric. Textual similarity is evaluated using BLEU (Papineni et al., 2002) and BERTScore (Zhang et al., 2019) against both the complex and reference texts.

4.5. Robust Inference

To mitigate the inherent non-determinism of LLMs, we did multiple independent inference runs (five for smaller models, three for larger models fine-tuned with LoRA (Hu et al., 2021)). The results were aggregated to report the mean and standard deviation for each metric. We reduced temperature to 0.1 to select the most probable token at each step.

5. Experimental Setup

5.1. Dataset Experiments

Stratified Partitioning. To ensure that the splits are representative of the overall dataset, we carry out stratified partitioning experiments. We evaluated multiple stratification strategies by sampling based on readability level metrics (FKGL, ARI, Dale-Chall) and length (word and char count) of the complex source text. To evaluate the preservation of the original distribution, we utilized the Kolmogorov–Smirnov (KS) goodness-of-fit test using `SciPy` package (Virtanen et al., 2020) to compare the splits against the full dataset, repeated over 10 random seeds for robustness. The winning approach yielding the lowest KS distance was chosen to generate the final 80/10/10 splits.

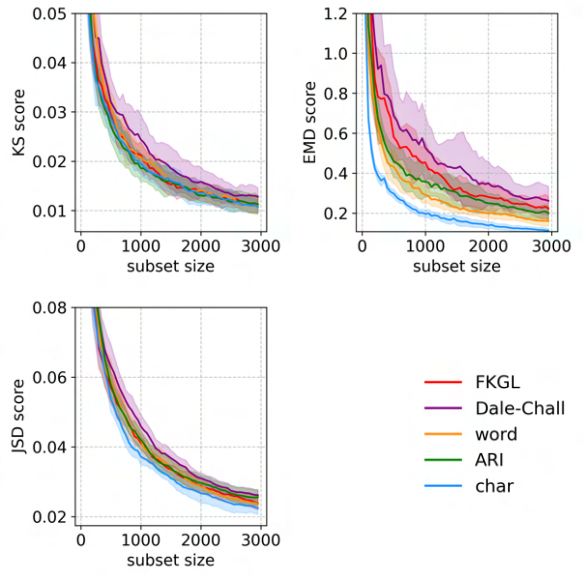


Figure 4: WIKILARGE. *Global* sampling with stratification by readability and length, measured in terms of KS, EMD and JSD. Stratification by N chars shows smallest divergence across all metrics.

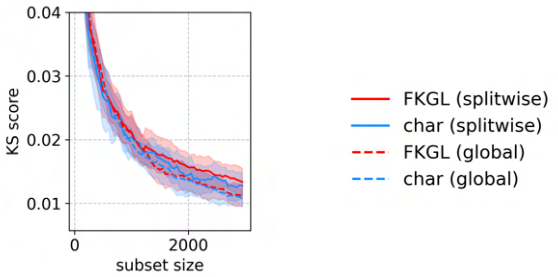


Figure 5: WIKILARGE. A comparison of split-based and split-agnostic sampling shows the latter yields lower divergence across both length and readability, with FKGL and N chars yielding similar results.

Sampling from WIKILARGE. Given that the WIKILARGE corpus contains a native train/dev/test split, we conduct a stratified sampling experiment to determine an optimal strategy to minimize divergence across all five control attributes. We compared two approaches: *global* sampling (from the entire dataset pool) and *split-wise* sampling (from the native train/dev/test partitions), exploring subset sizes in the 100–3,100 range (step=20) over 10 random seeds. The metrics used are the Kolmogorov-

SARI ↑		LENS ↑		MAE ↓	
1B	8B	1B	8B	1B	8B
MED-EASI FKGL					
43.68	*36.20*	*59.03*	*68.95*	*3.37*	*4.89*
42.64 -1.04	43.99 +7.77	57.33 -1.70	59.83 -9.12	3.96 +0.51	2.91 ?
MED-EASI char					
32.38	*36.76*	*21.23*	*62.89*	*1.23*	*0.37*
42.43 +10.05	45.25 +8.49	55.74 +34.51	59.90 -2.99	0.39 -0.84	0.21 -0.16
SIMPA FKGL					
42.90	*23.03*	*55.29*	*69.34*	*2.92*	*6.46*
42.15 -0.75	41.81 +18.78	55.05 -0.24	55.42 -13.92	2.56 -0.36	2.18 -4.28
SIMPA char					
33.07	*27.31*	*26.22*	*61.86*	*0.73*	*0.33*
38.95 +5.88	59.14 +31.83	53.93 +27.21	57.54 -4.32	0.18 -0.55	0.04 -0.29
WIKILARGE FKGL					
40.55	*25.69*	*50.55*	*81.37*	*3.66*	*2.69*
38.21 -2.34	37.73 +12.04	61.55 +11.00	60.97 -20.40	4.25 +0.59	2.94 +0.25
WIKILARGE char					
34.49	*37.54*	*22.53*	*66.60*	*1.60*	*0.33*
37.90 +3.41	38.94 +1.40	60.12 +37.59	61.55 -5.05	0.40 -1.20	0.24 0.07
NEWSELA FKGL					
28.52	*25.69*	*61.37*	*81.37*	*2.89*	*2.69*
37.60 +9.08	42.13 +16.44	40.34 -21.03	50.46 -30.91	2.51 -0.38	7.78 +5.09
NEWSELA char					
33.25	*26.88*	*44.51*	*70.74*	*0.35*	*0.56*
34.45 +1.25	38.79 +11.91	36.46 -8.05	49.46 -21.28	0.72 +0.37	0.36 -0.20

Table 2: IFT of 1B and 8B Llama models improves SARI and MAE over the non-fine-tuned baseline, with stronger gains for larger models and compression-based controls, whereas LENS does not consistently improve and in some cases decreases after fine-tuning.

Smirnov test (KS), Jensen–Shannon Divergence (JSD), and Earth Mover's Distance (EMD), computed with `SciPy` (Virtanen et al., 2020).

5.2. Instruction Fine-Tuning Experiments

Main Experiment. The primary experiment in this study evaluates the effectiveness of instruction fine-tuning (IFT) for CATS across four distinct domains and five control attributes. We fine-tuned our selection of LLMs on the full training sets of MED-EASI, SIMPA, WIKILARGE, and NEWSELA. Each model was trained separately for each of the five control attributes: readability (FKGL, ARI, and Dale-Chall) and length (character and word compression).

Mono-Datasets Ablation. We performed a fine-tuning experiment on the filtered subsets of the dataset described in Section 3.3. We hypothesized that removing uninformative instances might reduce computational demands and provide a clearer gradient for the model to learn the correlation between control tokens and linguistic outcomes.

Scaling Experiment. We investigated the relationship between model size and performance by scaling from 1B to 14B parameters within models of the same family and, if available, same model generation. The scaling experiment was conducted only on the MED-EASI and SIMPA datasets using FKGL and char compression as control attributes.

6. Results

6.1. Data Experiments

Stratified Partitioning. Using FKGL, ARI, Dale-Chall, character count, and word count as candidate stratification variables, we measured divergence with the KS statistic across multiple bin sizes and random seeds, in order to apply the same partitioning strategy to all datasets. Stratification by FKGL yielded lowest KS divergence between splits (see Fig. 4 and 5 for a comparison with sampling from WIKILARGE). The results led to our choice of FKGL as the primary stratification variable for split creation. This demonstrates that relying on native dataset splits or random partitioning can introduce substantial distributional mismatch, potentially confounding evaluation but also damaging fine-tuning.

Sampling from WIKILARGE. Across all bin sizes, *global* sampling resulted in smaller distribution divergence between the original dataset and its subsets. Fig. 4 and 5 show a nearly-monotonic decrease in the divergence score (KS) following subset size increase. Using N chars as stratification variable lead to lower divergence scores in both *global* and *split-wise* setups (Fig. 5).

6.2. Instruction Fine-Tuning

Main Experiment. Table 4 reports simplification quality (SARI, LENS) and controllability (MAE) across four datasets, three model families, and five control attributes. Performance appears to vary more by *dataset* and *attribute* than by model size.

Dataset effects. SIMPA yields the strongest results in terms of SARI across most model families and control attributes. This pattern does not hold for LENS, where MED-EASI and WIKILARGE sometimes higher scores depending on the model and control attribute. This divergence suggests that SARI and LENS reward different aspects of simplification quality, with their agreement varying greatly across models, control attributes and datasets (see Appendix C). NEWSELA (document-level) shows the lowest quality in terms of both SARI and LENS among all datasets, but not in terms of MAE. See model output examples in Table 1 and Appendix D.

Model-family trends. Across control attributes, Mistral-7B frequently achieves strong SARI scores on MED-EASI and SIMPA. This advantage is less consistent under LENS, where Qwen models often match or outperform while also achieving low control error (MAE). Qwen exhibits most stable performance across datasets and metrics, whereas Mistral and Llama show pronounced dataset-specific peaks. Increasing model size does not lead to consistent improvements: gains are non-monotonic

(a) SARI ↑

Llama				Qwen			
1B	3B	8B	13B	1.7B	4B	8B	14B
42.64	43.05	43.99	**52.02**	50.50	51.17	51.10	51.93
42.15	60.16	41.81	**67.78**	65.52	65.07	67.65	65.90

(b) LENS ↑

Llama				Qwen			
1B	3B	8B	13B	1.7B	4B	8B	14B
57.33	56.82	**59.90**	57.41	58.18	58.96	59.27	58.88
55.05	**59.13**	55.42	58.35	58.10	58.09	58.59	58.38

(c) MAE ↓

Llama				Qwen			
1B	3B	8B	13B	1.7B	4B	8B	14B
3.96	3.49	2.91	2.99	2.72	2.91	**2.34**	2.61
2.56	1.11	2.18	1.27	1.22	1.08	**1.03**	1.06

(d) SARI ↑

Llama				Qwen			
1B	3B	8B	13B	1.7B	4B	8B	14B
42.43	41.58	45.25	52.06	**52.09**	51.33	51.91	50.72
38.95	64.29	59.14	67.17	62.82	62.30	**65.96**	62.94

(e) LENS ↑

Llama				Qwen			
1B	3B	8B	13B	1.7B	4B	8B	14B
55.74	58.45	59.83	57.47	**60.54**	58.91	59.94	59.19
53.93	58.45	57.54	58.28	58.29	57.55	**58.49**	57.81

(f) MAE ↓

Llama				Qwen			
1B	3B	8B	13B	1.7B	4B	8B	14B
0.39	0.36	0.21	0.29	0.18	0.22	0.19	**0.15**
0.18	**0.04**	**0.04**	0.05	0.07	0.05	**0.04**	**0.04**

Table 3: Scaling experiment with a broader parameter size range. Rows in green represent results on MED-EASI, whereas rows in blue represent results on SIMPA. Top value per row and metric is in bold, second-best value is underscored. (a) - (c) FKGL, (d) - (f) char compression.

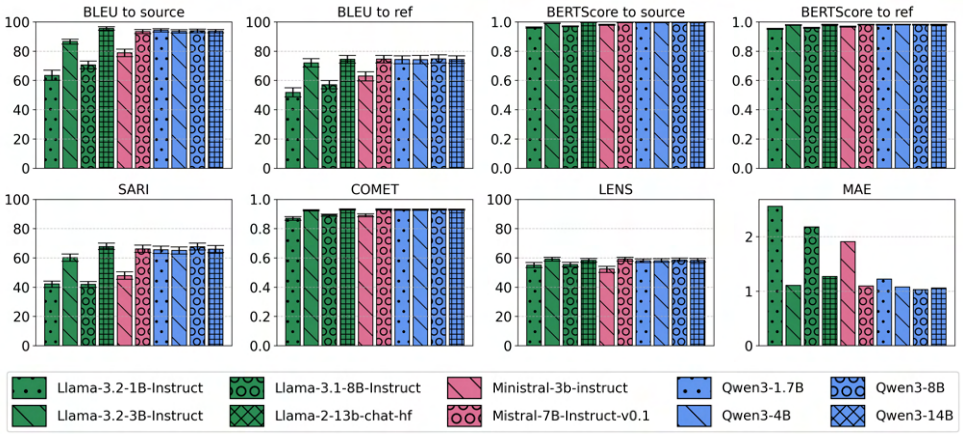


Figure 6: Dataset: SIMPA. Control attribute: FKGL. Scaling does not always boost performance. However, we observe strong positive correlation between SARI and COMET, and a strong negative correlation between SARI/COMET and error-based metrics across most datasets and control attributes.

and vary depending on the dataset, control attribute, and evaluation metric.

Baseline comparison. Under identical inference conditions, only Llama models consistently produce well-formed outputs suitable for evaluation (see Section 9). Comparing fine-tuned models to their non-fine-tuned counterparts reveals a consistent improvement in SARI and MAE, particularly for larger models and compression-based controls (see Table 2 and Appendix B). This trend does not extend to LENS, where fine-tuning leads to drops in LENS despite gains in SARI and MAE.

Mono-Datasets Ablation. Fine-tuning on mono-datasets did not lead to consistent improvements across metrics, but rather introduced a trade-off between simplification quality and controllability (Table 4). While SARI often decreases or remains comparable, LENS improves on MED-EASI and WIKILARGE. Controllability (MAE) does not consistently benefit from mono-dataset training. Qwen models show worse MAE, while Llama and Mistral models show minor fluctuations in both directions. Filtering for monotonic readability reduction seems to either damage training diversity, harmfully shifting the training distribution, or creates dataset that are too small for fine-tuning and evaluation.

Scaling experiment. Scaling does not lead to consistent improvements across evaluation metrics (Table 3, Appendix A). For FKGL control, SARI generally increases with model size, with larger checkpoints achieving the highest scores. LENS, in contrast, is less affected by model sizes change and remains stable especially across the Qwen family. Controllability (MAE) generally improves with scale. For char-compression control, the effect of scaling is even less pronounced. Scaling primarily benefits edit-based metrics such as SARI and, to some extent, MAE, while offering limited gains in LENS.

7. Discussion

Controllable simplification is not only a modeling problem, but is fundamentally constrained by *data and evaluation*. Performance differences are driven

Data	Ctrl.	SARI ↑						LENS ↑						MAE ↓					
		Llama		Mistral		Qwen		Llama		Mistral		Qwen		Llama		Mistral		Qwen	
		1B	8B	3B	7B	1.7B	8B	1B	8B	3B	7B	1.7B	8B	1B	8B	3B	7B	1.7B	8B
MED-EASI																			
base	ARI	42.47	43.47	44.78	**50.90**	50.32	50.55	56.89	**60.17**	50.20	59.28	58.74	59.33	4.39	3.35	8.19	3.30	3.63	**2.86**
base	DC	42.04	44.03	43.85	**52.18**	50.29	50.49	55.21	**59.64**	50.89	57.32	57.14	57.62	1.54	1.27	1.66	1.29	1.39	**1.16**
base	char	42.43	45.25	44.86	52.03	**52.09**	51.91	55.74	59.83	51.25	57.42	**60.54**	59.94	0.39	0.21	0.70	0.28	**0.18**	0.19
base	word	42.59	45.31	44.68	51.23	51.13	**52.00**	56.75	59.97	51.78	57.51	**60.62**	59.81	0.37	**0.24**	0.75	0.29	**0.24**	0.26
base	FKGL	42.64	43.99	43.05	**51.87**	50.50	51.10	57.33	**59.90**	–	58.28	58.18	59.27	3.96	2.91	3.49	2.85	2.72	**2.34**
mono	FKGL	41.95	43.43	44.26	**50.34**	47.81	49.53	59.27	**61.65**	50.09	56.45	58.34	59.26	3.61	**2.71**	4.12	3.67	3.44	2.93
SIMPA																			
base	ARI	41.47	40.64	52.13	64.48	65.17	**65.77**	55.59	53.83	54.25	**58.85**	58.44	58.48	3.05	2.45	2.53	1.24	**1.21**	1.25
base	DC	41.80	42.87	47.75	**65.75**	65.30	65.72	55.17	55.79	51.62	**58.49**	57.56	58.36	0.81	0.68	0.78	0.53	0.51	**0.47**
base	char	38.95	59.14	50.84	65.56	62.82	**65.96**	53.93	57.54	52.67	**58.68**	58.29	58.49	0.18	**0.04**	0.23	0.05	0.07	**0.04**
base	word	40.42	58.49	54.23	62.52	59.09	**66.50**	53.49	56.80	54.49	57.44	55.82	**58.16**	0.13	0.05	0.06	**0.04**	**0.04**	0.08
base	FKGL	42.15	41.81	47.87	66.10	65.52	**67.65**	55.05	55.42	52.35	**59.10**	58.10	58.59	2.56	2.18	1.91	1.10	1.22	**1.03**
mono	FKGL	45.32	45.31	50.78	57.44	57.26	**58.21**	51.07	**58.06**	53.99	56.92	57.74	57.53	2.94	2.70	2.86	2.79	**2.34**	2.57
WIKILARGE																			
base	ARI	38.34	37.78	37.89	48.18	49.45	**63.10**	60.64	**61.88**	48.95	57.20	55.24	56.87	4.69	3.78	16.87	2.81	2.90	**1.44**
base	DC	38.10	37.85	40.23	48.86	48.71	**49.30**	60.91	**62.18**	50.02	54.74	54.44	56.65	2.33	2.40	2.03	1.61	**1.56**	1.63
base	char	37.90	38.94	39.25	48.28	50.92	**51.67**	**60.12**	38.94	52.01	57.49	53.97	55.92	0.39	0.24	0.34	0.24	0.14	**0.12**
base	word	37.65	40.79	39.24	50.10	50.49	**51.41**	60.10	**61.14**	52.58	55.77	54.76	55.00	0.44	0.20	0.45	0.22	0.17	**0.13**
base	FKGL	38.21	37.73	39.10	49.11	**49.47**	48.66	**61.55**	60.97	46.89	57.56	54.60	56.28	4.25	2.94	8.24	2.39	2.50	**2.24**
mono	FKGL	40.38	41.38	39.64	44.59	**44.77**	44.08	**65.54**	63.40	50.18	59.17	60.26	59.77	3.46	**2.89**	7.87	3.41	3.43	3.08
NEWSELA																			
base	FKGL	37.60	42.13	31.02	43.41	**43.51**	43.34	40.34	50.46	19.98	47.53	48.00	49.62	2.51	7.78	26.46	2.77	2.43	**2.35**
base	char	34.45	38.79	31.49	41.62	43.09	**43.52**	36.46	49.46	20.97	47.61	46.97	**49.82**	0.72	0.36	1.39	0.31	**0.26**	0.27

Table 4: Main experiment results show strong performance of the fine-tuned Qwen models closely followed by the larger Mistral. "DC" stands for Dale-Chall. *"Mono"* refers to the monotonically filtered subsets of the respective dataset; *"base"* refers to its full version. "Word" and "char" denote word- and character-level compression ratio, respectively. Higher is better for SARI and LENS; lower is better for MAE.

by signal availability (attribute variation, distribution match) and by how we measure control. Building CATS systems is largely an exercise in curating *controllable signal* and *measuring target compliance*, with model choice and scale playing a secondary role provided a suitable fine-tuning framework.

Splits and sampling shape training and evaluation. Our stratified partitioning experiments (Section 6.1) show that random or native splits risk distributional mismatch in control attributes and potentially confound training and evaluation. Practically, controllable ATS benchmarks should report split creation and verify representativeness with divergence checks before attributing gains to modeling.

Scale is not a reliable proxy for controllability. We observe non-monotonic gains with increasing model size (Table 3, Fig. 6). Smaller models can be competitive: targeted IFT and data properties dominate raw scale; comparing models without controlling for data signal can prompt wrong conclusions.

Readability control is learnable, compression control is insufficient in sentence-level datasets. Across datasets, readability targets (FKGL/ARI/Dale-Chall) are learned more consistently than length. Our attribute distribution analysis indicates that compression targets provide weak training signal because many sentence-aligned corpora contain minimal complex-simple length variation (Fig. 1 and 2). Progress on compression-controllable ATS requires dedicated datasets that explicitly encode diverse compression ratios.

Excessive data cleaning risks drowning signal. The mono-datasets ablation shows slightly degraded performance, suggesting that strict filtering might remove diversity and shift the training distribution. Retaining broad coverage of attribute values may be better than enforcing monotonic readability reductions, especially with smaller datasets.

Measuring control requires integration of error-based metrics. Traditional metrics (SARI, LENS, BLEU) do not quantify alignment to the target and can reward copying behavior. Dedicated error-based measures are key for measure of control compliance and are necessary to evaluate CATS systems when the objective is *target matching*. While we observe strong (negative) correlation between SARI/LENS and MAE (see Fig. 6), a holistic approach to CATS evaluation requires multiple dimensions of simplification, including deviation from the target value (controllability), fluency (grammaticality) and meaning preservation (adequacy).

Robustness is part of evaluation. Because LLM outputs vary with decoding and random seeds, multi-seed evaluation is essential for stable comparisons. Our inference protocol prevents over-interpreting single-run results.

8. Conclusion

We investigated the efficacy of instruction fine-tuning with discrete control tokens to steer open-source LLMs toward readability and compression targets. Our experiments demonstrated that IFT with discrete control tokens is a lightweight and flexible method to transform open-source LLMs into steerable simplification systems. While we observe a positive correlation between model size and performance improvement, some outliers (Qwen3-1.7B) match or outperform larger counterparts.

Our data experiments across common text simplification datasets reveal a crucial limitation: the richness of control-attribute signal in the training

data limits how well the model can learn to perform an attribute-specific simplification. With most sentence-level datasets showing minimal compression in the complex-simple pairs, the model largely fails to learn to compress, whereas a pronounced difference in the FKGL value in the complex-simple pairs allows the model to learn to generate predictions approximating a target readability level.

We urge for a thoughtful selection of the stratification variable, in particular in a multi-control-attribute setup. As we demonstrated in the sampling and partitioning experiments, defaulting to the dataset's native splits may lead to an unwelcome divergencies in the attribute distribution between the dataset and its subsets. Running sampling by several control attributes before picking one yielding the lowest divergence provides an effective safeguard.

9. Limitations

Automatic evaluation and metric validity. We rely on automatic metrics for (i) simplification quality (SARI, LENS, COMET), (ii) similarity (BLEU, BERTScore), and (iii) controllability (MAE on target attributes). These metrics only partially capture human notions of simplicity, adequacy, and fluency: e.g., SARI is biased toward lexical edits, similarity metrics reward copying, LENS is designed specifically to align text simplification quality with human evaluation, but its agreement with other metrics ranges widely between datasets and control attributes (see Appendix C). While COMET is not designed for text simplification, MAE quantifies attribute matching but does not guarantee that outputs are acceptable simplifications: human evaluation would be indispensable to validate whether lower attribute error corresponds to better perceived controllability and readability.

Fine-tuning techniques. By using LoRA fine-tuning for models above 3B size, we effectively adopt two distinct fine-tuning approaches for smaller and larger LLMs. This puts a limitation on the comparability of model performance evaluation in the scaling analysis. We also adopt two sets of fine-tuning configurations (with and without LoRA) arrived at through hyperparameter tuning.

Baseline comparison. We are unable to provide a fully uniform non-fine-tuned baseline across Llama, Qwen, and Mistral. Our IFT pipeline relies on a complex shared prompting format (model-native chat template + dynamically inserted control tokens). Llama models largely follow instructions and produce coherent outputs, but the Qwen and Mistral lineup often seem to be overwhelmed by the complex prompt template and tend to produce degenerate output (empty or malformed output, prompt repetition, output repetition loops), preventing meaningful comparison and automatic evaluation. Using simpler or model-specific prompts could yield stronger prompt-only baselines, but would not be directly comparable to the IFT setting due to the altered conditioning format and effective task definition. Allowing model-specific baseline prompts would risk conflating controllability with prompt engineering.

Representativeness of sampled data. We use subsets for WIKILARGE and NEWSELA to keep experiments tractable. While we minimize *distributional divergence* between the original corpora and our subsets using stratified sampling (measured via KS/JSD/EMD), some mismatch in attribute distributions can remain, especially in the tails. As a result, reported controllability and simplification scores may differ when training/evaluating on the full datasets.

Cross-attribute analysis. We cannot draw robust comparative observations about controllability among different control attributes, because they have different scales. To make such comparisons possible, it would be necessary to normalize the attribute values and attribute-specific errors.

Language inclusiveness. We work exclusively with English-language datasets, which naturally limits the generalization of our findings. Scarcity of expert-generated, well-aligned simplification corpora is even more pronounced in other languages.

10. Acknowledgments

This work was supported by the Swiss Innovation Agency Innosuisse, Flagship Inclusive Information and Communication Technology (IICT), funding no. PFFS-21-47. We sincerely thank Prof. Sarah Ebling for her valuable contributions to the study.

11. Plain Language Summary

Some texts are difficult to read and understand for reasons including age, literacy, or language mastery; texts may also be difficult if they come from a specific knowledge field unfamiliar to their reader, such as medical documents or government decisions. The goal of automatic text simplification is to make texts more accessible. Ideally, such systems should also give their users the power to control how simple they want the text to be in terms of its length, vocabulary and syntactic complexity.

In this work, we explore a method that teaches Large Language Models how to simplify text to a desired reading level or length by showing them

examples of good simplifications. We apply this method to texts from the fields of medicine, public administration, news and encyclopedic knowledge. We find that models can learn to control readability reasonably well, which can be measured by comparing predefined gold standard simplification with the text produced by the model and measuring how different they are. Controlling text length is a harder task: because complex and simple texts in the existing datasets tend to have similar length, there is not much length transformation the model can learn to imitate.

We also show that results depend strongly on how the data is selected, prepared and how performance is measured. Common evaluation metrics do not fully capture whether the model follows the instructions. Our findings suggest that the future of controllable simplification systems depends both on improving model capabilities and on improving data quality and evaluation methods.

12. Bibliographical References

Sweta Agrawal and Marine Carpuat. 2024. Do Text Simplification Systems Preserve Meaning? A Human Evaluation via Reading Comprehension. *Transactions of the Association for Computational Linguistics*, 12:432–448. Place: Cambridge, MA Publisher: MIT Press.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020. Data-Driven Sentence Simplification: Survey and Benchmark. *Computational Linguistics*, 46(1):135–187. Place: Cambridge, MA Publisher: MIT Press.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. The (Un)Suitability of Automatic Evaluation Metrics for Text Simplification. *Computational Linguistics*, 47(4):861–889. Place: Cambridge, MA Publisher: MIT Press.

Chandrayee Basu, Rosni Vasu, Michihiro Yasunaga, and Qian Yang. 2023. Med-EASi: Finely Annotated Dataset and Models for Controllable Simplification of Medical Texts. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12):14093–14101.

Jeanne Sternlicht Chall and Edgar Dale. 1995. Readability Revisited : the New Dale-Chall Readability Rormula.

Raman Chandrasekar and Srinivas Bangalore. 1997. Automatic Induction of Rules for Text Simplification. *Knowl. Based Syst.*, 10:183–190.

Edgar Dale and Jeanne Sternlicht Chall. 1948. A Formula for Predicting Readability.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The Llama 3 Herd of Models. *arXiv e-prints*, pages arXiv–2407.

Isabel Espinosa-Zaragoza, José Abreu-Salas, Elena Lloret, Paloma Moreda, and Manuel Palomar. 2023. A Review of Research-Based Automatic Text Simplification Tools. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 321–330, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Natalia Grabar and Horacio Saggion. 2022. Evaluation of Automatic Text Simplification: Where Are We Now, Where Should We Go from Here. In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, pages 453–463, Avignon, France. ATALA.

J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *ArXiv*, abs/2106.09685.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Tannon Kew, Alison Chi, Laura Vásquez-Rodríguez, Sweta Agrawal, Dennis Aumiller, Fernando Alva-Manchego, and Matthew Shardlow. 2023. BLESS: Benchmarking Large Language Models on Sentence Simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13291–13309, Singapore. Association for Computational Linguistics.

Tannon Kew and Sarah Ebling. 2022. Target-Level Sentence Simplification as Controlled Paraphrasing. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 28–42, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Peter Kincaid, Robert P. Fishburne, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel.

Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2023. LENS: A learnable evaluation metric for text simplification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16383–16408, Toronto, Canada. Association for Computational Linguistics.

Louis Martin, Éric de la Clergerie, Benoît Sagot, and Antoine Bordes. 2020. Controllable Sentence Simplification. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4689–4698, Marseille, France. European Language Resources Association.

Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. 2022. MUSS: Multilingual Unsupervised Sentence Simplification by Mining Paraphrases. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1651–1664, Marseille, France. European Language Resources Association.

Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P. Dinu. 2017. Exploring Neural Text Simplification Models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 85–91, Vancouver, Canada. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Carolina Scarton, Gustavo Paetzold, and Lucia Specia. 2018. SimPA: A Sentence-Level Simplification Corpus for the Public Administration Domain. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Matthew Shardlow. 2014. A Survey of Automated Text Simplification. *International Journal of Advanced Computer Science and Applications*, 4.

Kim Cheng Sheang and Horacio Saggion. 2021. Controllable Sentence Simplification with a Unified Text-to-Text Transfer Transformer. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 341–352, Aberdeen, Scotland, UK. Association for Computational Linguistics.

E A Smith and R. Senter. 1967. Automated Readability Index. *AMRL-TR. Aerospace Medical Research Laboratories*, pages 1–14.

Nicolas Spring, Annette Rios, and Sarah Ebling. 2021. Exploring German Multi-Level Text Simplification. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1339–1349, Held Online. INCOMA Ltd.

Hieu Tran, Zonghai Yao, Lingxi Li, and Hong Yu. 2025. ReadCtrl: Personalizing Text Generation with Readability-Controlled Instruction Learning. In *Proceedings of the Fourth Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2025)*, pages 19–36, Albuquerque, New Mexico, US. Association for Computational Linguistics.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python.

Laura Vásquez-Rodríguez, Matthew Shardlow, Piotr Przybyła, and Sophia Ananiadou. 2021. Investigating Text Simplification Evaluation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 876–882, Online. Association for Computational Linguistics.

Tong Wang, Ping Chen, Kevin Amaral, and Jipeng Qiang. 2016a. An Experimental Study of LSTM Encoder-Decoder Model for Text Simplification. *arXiv preprint arXiv:1609.03663*.

Tong Wang, Ping Chen, John Rochford, and Jipeng Qiang. 2016b. Text Simplification using Neural Machine Translation. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 4270–7271. AAAI Press.

Sander Wubben, Antal van den Bosch, and Emiel Krahmer. 2012. Sentence Simplification by Monolingual Machine Translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1015–1024, Jeju Island, Korea. Association for Computational Linguistics.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in Current Text Simplification Research: New Data Can Help. *Transactions of the Association for Computational Linguistics*, 3:283–297. Place: Cambridge, MA Publisher: MIT Press.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing Statistical Machine Translation for Text Simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415. Place: Cambridge, MA Publisher: MIT Press.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*.

Kevin Yang and Dan Klein. 2021. FUDGE: Controlled Text Generation With Future Discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3511–3535, Online. Association for Computational Linguistics.

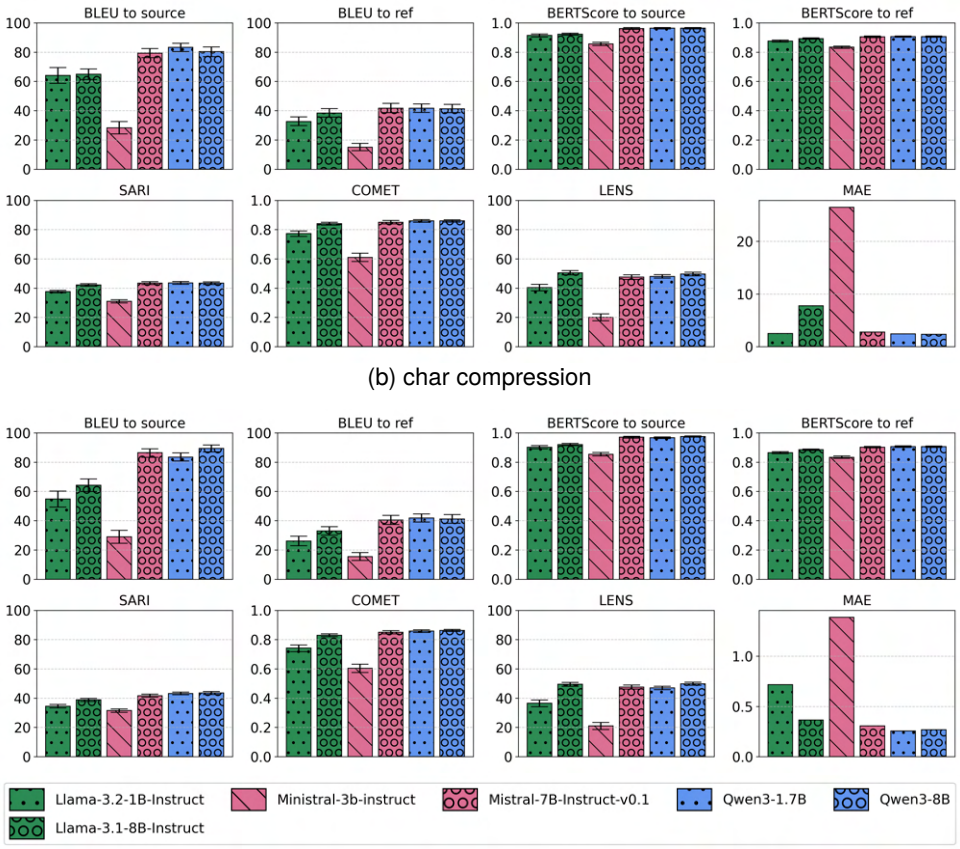
Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT. *ArXiv*, abs/1904.09675.

Xingxing Zhang and Mirella Lapata. 2017. Sentence Simplification with Deep Reinforcement Learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 584–594, Copenhagen, Denmark. Association for Computational Linguistics.

A. Model Performance Comparison

Figure 7: Model performance on NEWSELA.

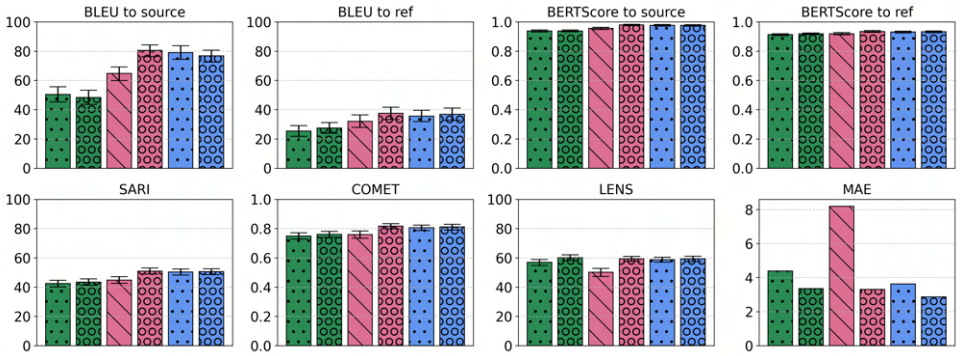
(a) FKGL



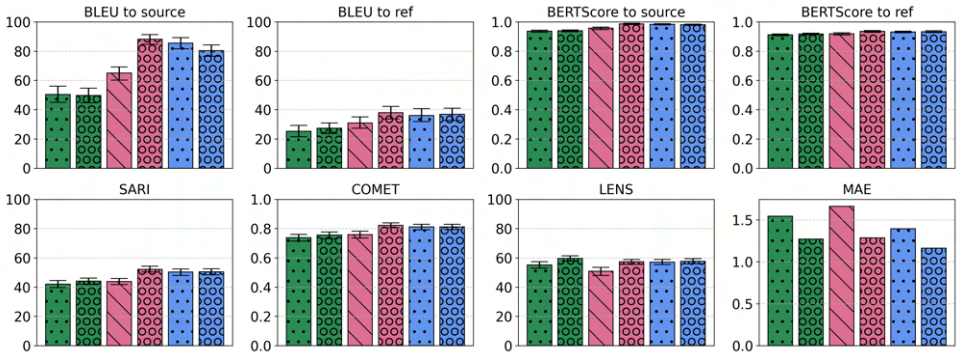
(b) char compression

Figure 8: Model performance on MED-EASI.

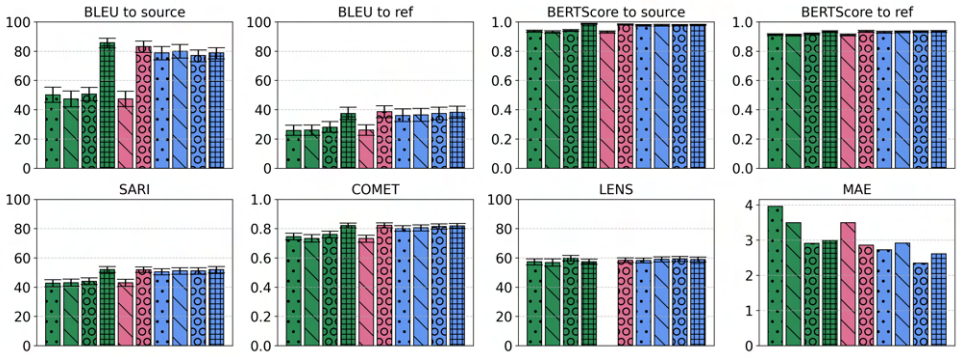
(a) ARI



(b) Dale-Chall



(c) FKGL



(d) char compression

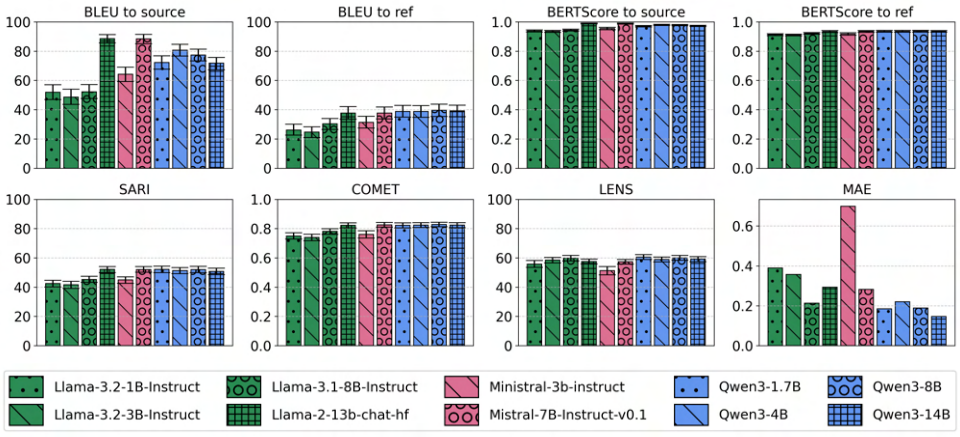
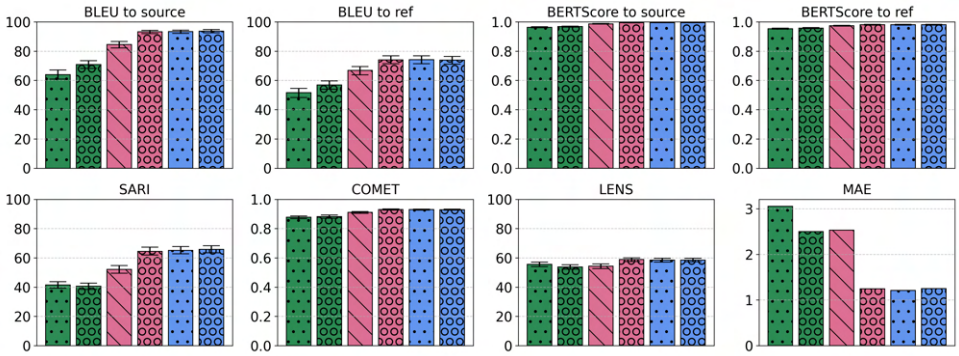
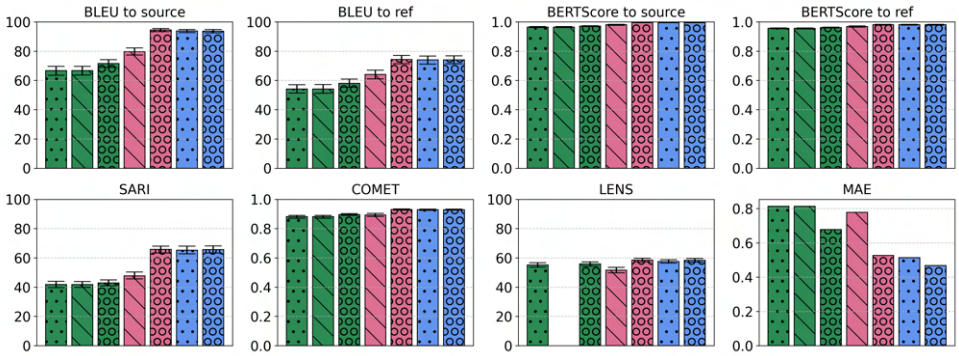


Figure 9: Model performance on SIMPA.

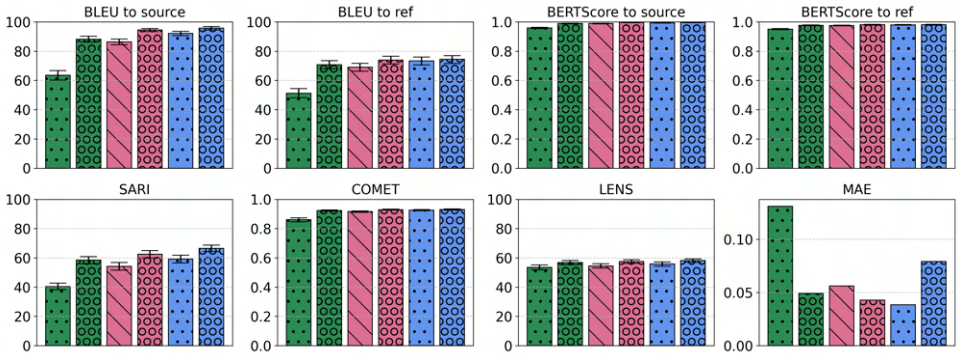
(a) ARI



(b) Dale-Chall



(c) word compression



(d) char compression

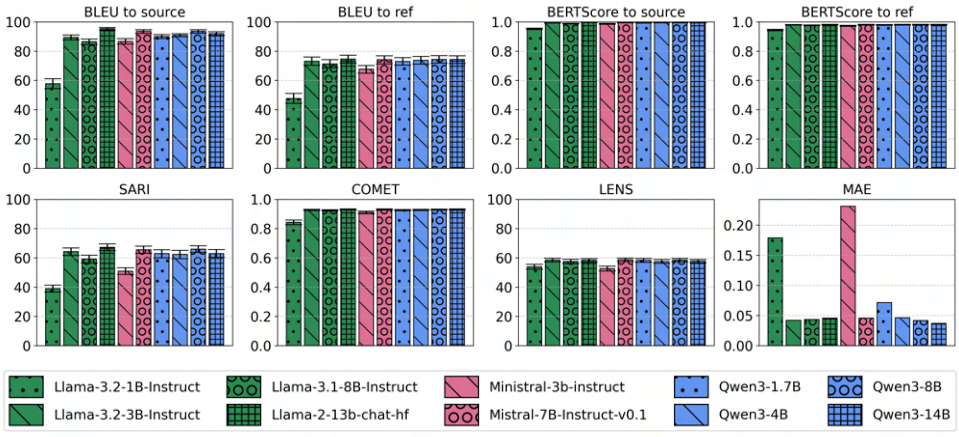
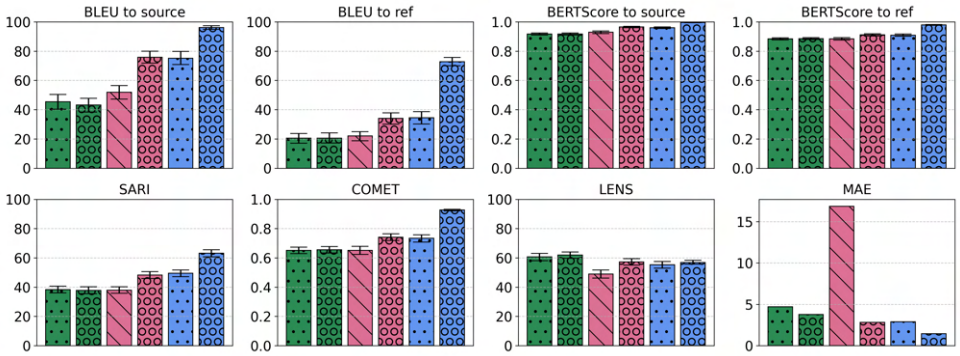
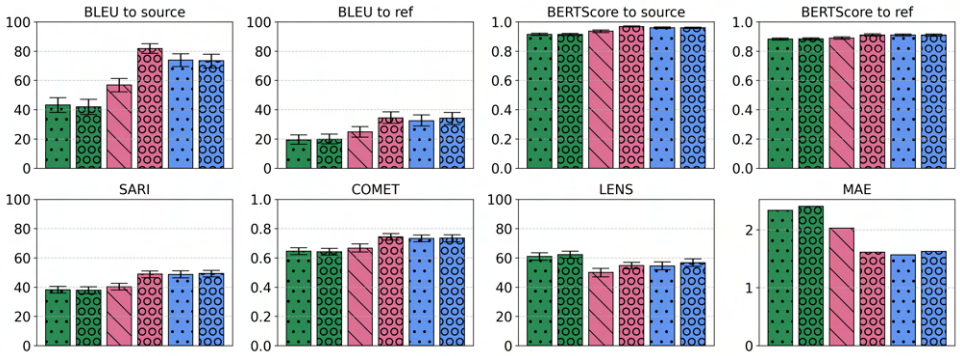


Figure 10: Model performance on WikiLarge.

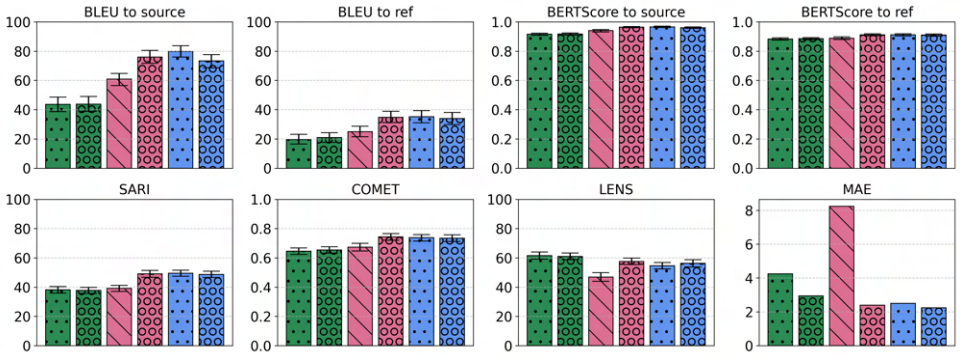
(a) ARI



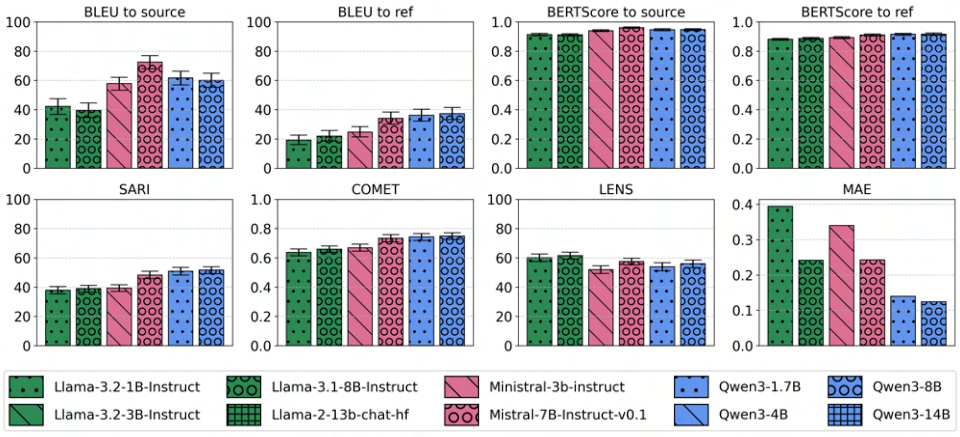
(b) Dale-Chall



(c) FKGL



(d) char compression



B. Baseline Controllability Analysis

Figure 11: Dataset: SIMPA

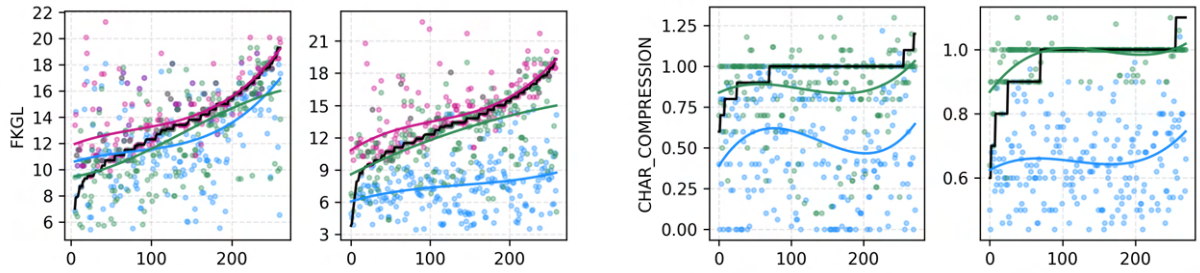


Figure 12: Dataset: MEDEASI

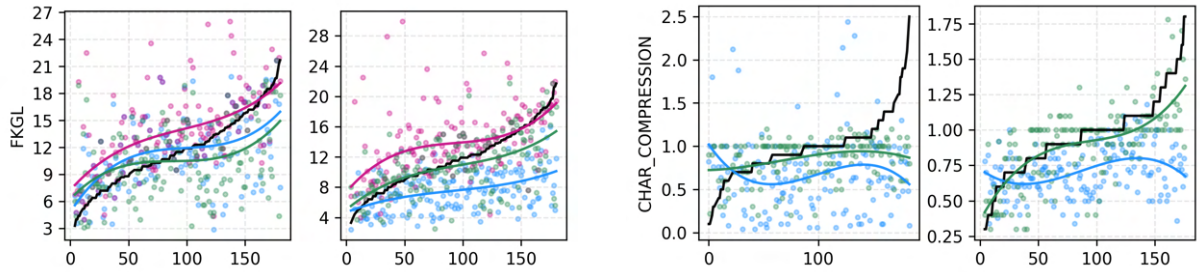


Figure 13: Dataset: WIKILARGE

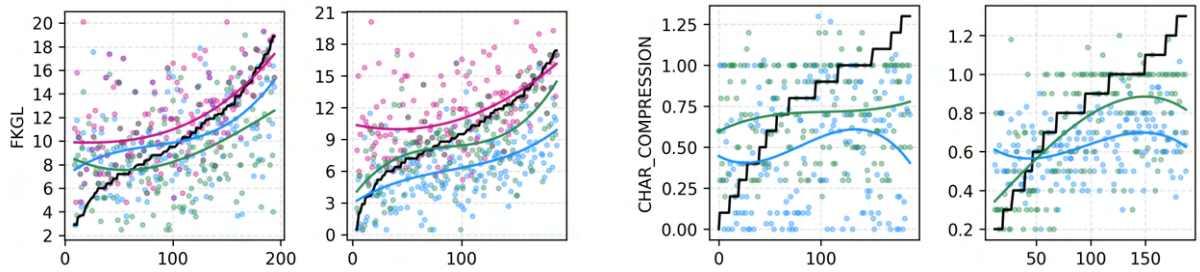


Figure 14: Dataset: NEWSELA

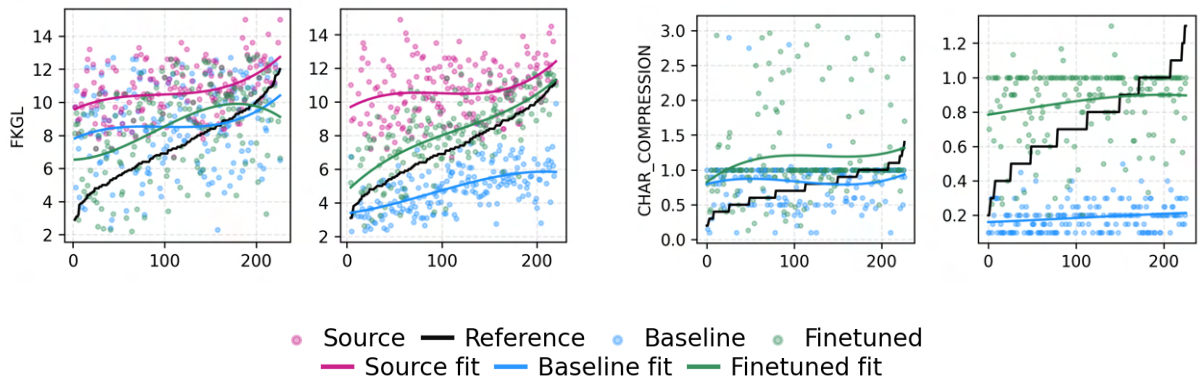


Figure 15: Comparison of controllability in terms of MAE between the instruction fine-tuned and non-fine-tuned models. For each control attribute: left plot is Llama-3.2-1B-Instruct, right is Llama-3.1-8B-Instruct.

C. Metric Correlation Analysis

Figure 16: Pearson correlation heatmaps analyzing metric correlation, aggregated across all models. Dataset: NEWSELA.

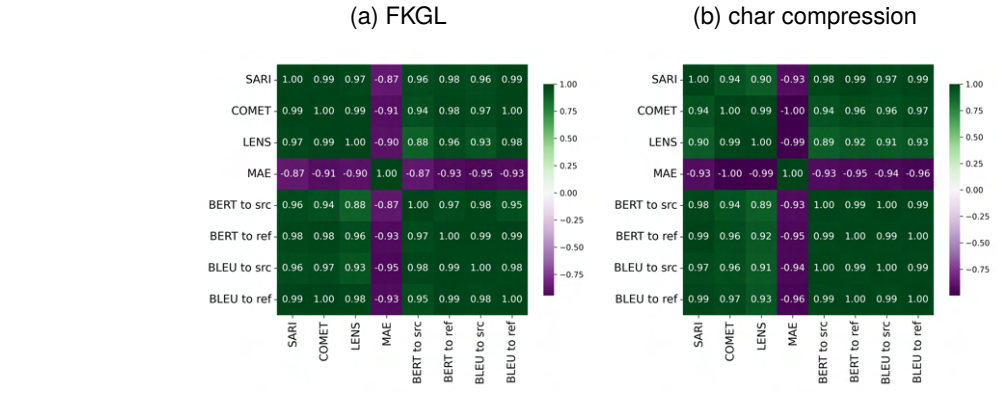


Figure 17: Pearson correlation heatmaps analyzing metric correlation, aggregated across all models. Dataset: MED-EASI.

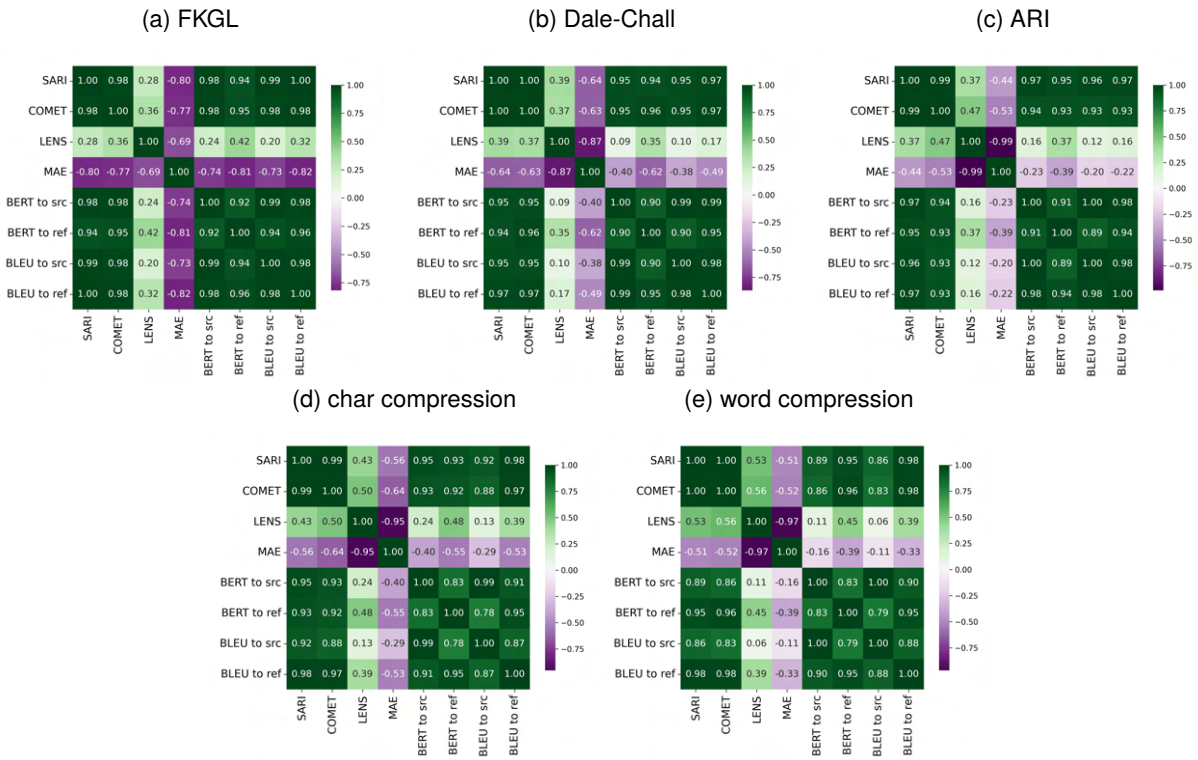
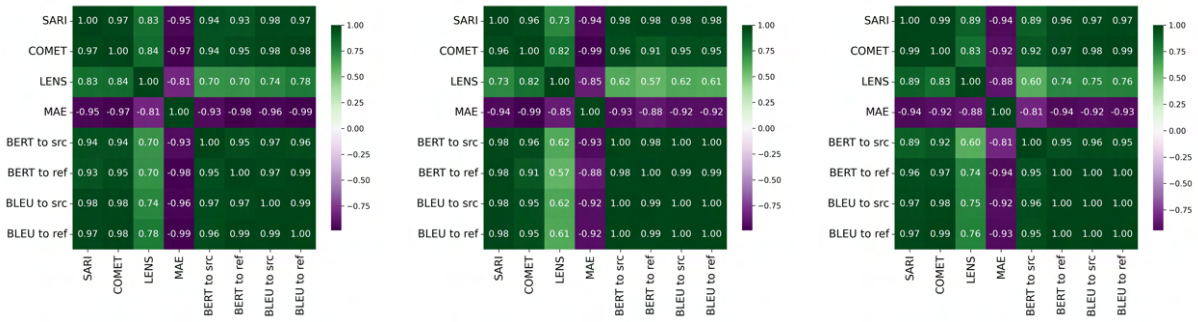


Figure 18: Pearson correlation heatmaps analyzing metric correlation, aggregated across all models. Dataset: SIMPA.

(a) FKGL (b) Dale-Chall (c) ARI



(d) char compression (e) word compression

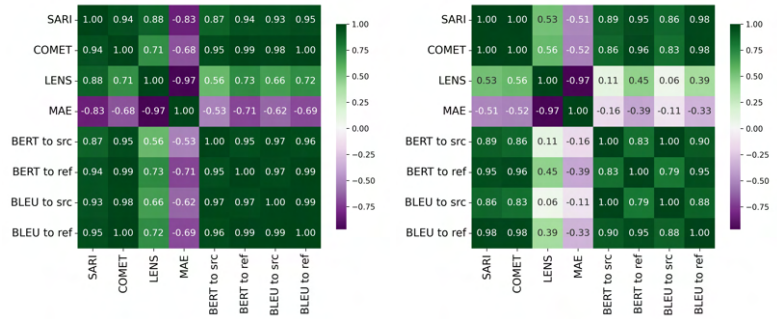
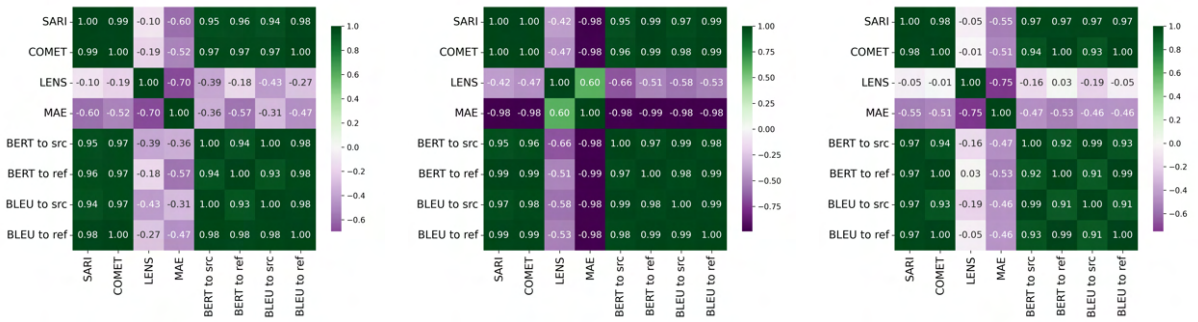
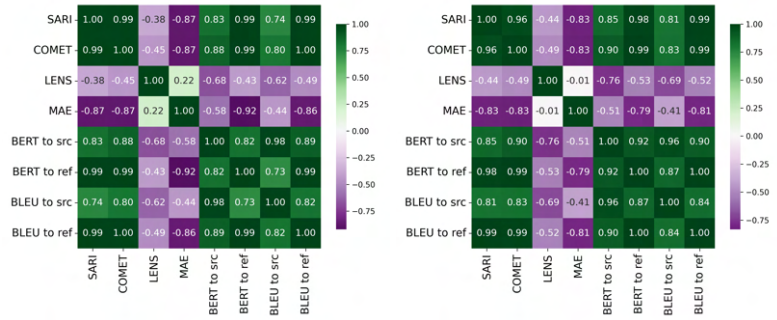


Figure 19: Pearson correlation heatmaps analyzing metric correlation, aggregated across all models. Dataset: WIKILARGE.

(a) FKGL (b) Dale-Chall (c) ARI



(d) char compression (e) word compression



D. Model Output Examples

Table 5: Dataset: WIKILARGE. Control Attribute: Dale-Chall. Instance ID: wikilarge_ori_252807. In this example, we see how different models focus on different information in the source text, resulting in very different simplifications. Further, reference simplification is much shorter than the source text and omits essential information. This is an example of a poor reference text, with lower adequacy and fluency than the automatically generated simplifications by various models.

	DC	**Text**
Source	9.5	"In public saunas it is more common that men and women go to the sauna separately, although people of both sexes may sometimes bathe together, for example in student clubs. Still, saunas are not associated with sex and sexuality".
Reference	**5.7**	"Some saunas are only for men , or only for women".
Llama-3.2-1B-Instruct	11.1	"The sauna is not associated with sex and sexuality".
Llama-3.1-8B-Instruct	7.3	"The sauna is a place where people go to relax".
Ministral-3b-instruct	9.3	"In public saunas it is more common that men and women go to the sauna separately, although people of both sexes may occasionally bathe together".
Mistral-7B-Instruct-v0.1	9.8	"In public saunas it is more common that men and women go to the sauna separately, although people of both sexes may sometimes bathe together, for example in student clubs".
Qwen3-1.7B	11.9	"Saunas are not associated with sex and sexuality".
Qwen3-8B	8.4	"In public saunas it is more common that men and women go to the sauna separately".

Table 6: Dataset: MED-EASI. Control Attribute: ARI. Instance ID: medeasi_000434. Simplifications generated by the model are very similar, further highlighting the effect of sentence splitting on the ARI score. Llama-3.2-1B-Instruct and Ministral-3b-instruct split the long sentence into two and drastically reduce the ARI score without any further transformations.

	ARI	**Text**
Source	20.1	"Symptoms include a sense of fullness, heaviness, aching, cramps, pain, tiredness, and paresthesias in the legs; these symptoms worsen with standing or walking and are relieved by rest and elevation".
Reference	**15.2**	"The swelling may not cause any symptoms, but some people feel fullness, heaviness, aching, cramps, pain, tiredness, and tingling in the legs".
Llama-3.2-1B-Instruct	11.8	"The symptoms include a sense of fullness, heaviness, aching, cramps, pain, tiredness, and numbness in the legs. These symptoms worsen with standing or walking and are relieved by rest and elevation".
Llama-3.1-8B-Instruct	19.6	"The symptoms include a sense of fullness, heaviness, aching, cramps, pain, tiredness, and numbness in the legs; these symptoms worsen with standing or walking and are relieved by rest and elevation".
Ministral-3b-instruct	10.7	"Symptoms include a sense of fullness, heavyness, aching, cramps, and tiredness. They worsen with standing or walking and are relieved by rest and elevation".
Mistral-7B-Instruct-v0.1	19.5	"Symptoms include a sense of fullness, heaviness, aching, cramps, pain, tiredness, and tingling in the legs; these symptoms worsen with standing or walking and are relieved by rest and elevation".
Qwen3-1.7B	20.1	"Symptoms include a sense of fullness, heaviness, aching, cramps, pain, tiredness, and paresthesias in the legs; these symptoms worsen with standing or walking and are relieved by rest and elevation".
Qwen3-8B	19.5	"Symptoms include a sense of fullness, heaviness, aching, cramps, pain, tiredness, and tingling in the legs; these symptoms worsen with standing or walking and are relieved by rest and elevation".

E. Prompt Template

Listing 1: Control-token-specific instructions dynamically inserted into the prompt.

```
"ARI":
  {"prompt": "INSTRUCTION: Simplify the following text such that its Automated Readability Index (ARI)
   score is approximately equal to that specified in the control token prepended to your generated
   simplification. The control token has the following format: <METRIC=VALUE>. The control token prepended
    to the source text indicates the ARI value of the source text. \nSOURCE TEXT: <ARI={SOURCE_VALUE}> {
   TEXT}\nEXPLANATION: {EXPLANATION}"}
"FKGL":
  {"prompt": "INSTRUCTION: Simplify the following text such that its Flesch-Kincaid Grade Level (FKGL)
   score is approximately equal to that specified in the control token prepended to your generated
   simplification. The control token has the following format: <METRIC=VALUE>. The control token prepended
    to the source text indicates the FKGL value of the source text. \nSOURCE TEXT: <FKGL={SOURCE_VALUE}> {
   TEXT}\nEXPLANATION: {EXPLANATION}"}
"DALE-CHALL":
  {"prompt": "INSTRUCTION: Simplify the following text such that its Dale-Chall Readability Score is
   approximately equal to that specified in the control token prepended to your generated simplification.
   The control token has the following format: <METRIC=VALUE>. The control token prepended to the source
   text indicates the Dale-Chall value of the source text. \nSOURCE TEXT: <DALE-CHALL={SOURCE_VALUE}> {
   TEXT}\nEXPLANATION: {EXPLANATION}"}
"CHAR_COMPRESSION":
  {"prompt": "INSTRUCTION: Simplify the following text such that the length of the output text (number of
   characters) relative to the source text is approximately equal to the ratio specified in the control
   token prepended to your generated simplification. The control token has the following format: <METRIC=
   VALUE>. \nSOURCE TEXT: {TEXT}\nEXPLANATION: {EXPLANATION}"}
"WORD_COMPRESSION":
  {"prompt": "INSTRUCTION: Simplify the following text such that the length of the output text (number of
   words) relative to the source text is approximately equal to the ratio specified in the control token
   prepended to your generated simplification. The control token has the following format: <METRIC=VALUE>.
    \nSOURCE TEXT: {TEXT}\nEXPLANATION: {EXPLANATION}"}
```

Listing 2: Control-token-specific explanations dynamically inserted into the prompt.

```
  "ARI": {
      "token": "<ARI={ARI_VALUE}>",
      "description": "Automated Readability Index",
      "value type": "float",
      "explanation": "The <ARI={TARGET_VALUE}> token specifies that the target Automated Readability Index
    (ARI) score should be approximately {TARGET_VALUE}. Lower values indicate simpler text."
  },
  "FKGL": {
      "token": "<FKGL={FKGL_VALUE}>",
      "description": "Flesch-Kincaid Grade Level",
      "value type": "float",
      "explanation": "The <FKGL={TARGET_VALUE}> token specifies that the target Flesch-Kincaid Grade Level
    should be approximately {TARGET_VALUE}. Lower values indicate simpler text."
  },
  "DALE-CHALL": {
      "token": "<DALE-CHALL={DALE_CHALL_VALUE}>",
      "description": "Dale-Chall Readability Score",
      "value type": "float",
      "explanation": "The <DALE-CHALL={TARGET_VALUE}> token specifies that the target Dale-Chall
   readability score should be approximately {TARGET_VALUE}. Lower values indicate simpler text."
  },
  "CHAR_COMPRESSION": {
      "token": "<CHAR_COMPRESSION={CHAR_COMPRESSION_VALUE}>",
      "description": "Character-level Compression Rate",
      "value type": "float",
      "explanation": "The <CHAR_COMPRESSION={TARGET_VALUE}> token specifies that the character-level
   compression rate should be around {TARGET_VALUE}. Values less than 1 indicate text reduction. Values
   greater than 1 indicate text expansion."
  },
  "WORD_COMPRESSION": {
      "token": "<WORD_COMPRESSION={WORD_COMPRESSION_VALUE}>",
      "description": "Word-level Compression Rate",
      "value type": "float",
      "explanation": "The <WORD_COMPRESSION={TARGET_VALUE}> token specifies that the word-level
   compression rate should be around {TARGET_VALUE}. Values less than 1 indicate text reduction. Values
   greater than 1 indicate text expansion."
  }
```

Figure 20: Formatted prompt examples.

(a) `Llama-3.2-1B-Instruct`

```
<|begin_of_text|><|start_header_id|>system<|
    end_header_id|>

Cutting Knowledge Date: December 2023
Today Date: 19 May 2025

You are a helpful expert in text simplification. You
    generate a simplified version of the text input
    by the user. You simplify the text according to
    the instructions given by the user. When asked
    to simplify a text, generate only the requested
    simplification, without any additional comments,
     notes or explanations.<|eot_id|><|
    start_header_id|>user<|end_header_id|>

INSTRUCTION: Simplify the following text such that
    its Flesch-Kincaid Grade Level (FKGL) score is
    approximately equal to that specified in the
    control token prepended to your generated
    simplification. The control token has the
    following format: <METRIC=VALUE>. The control
    token prepended to the source text indicates the
     FKGL value of the source text.
SOURCE TEXT: <FKGL=4.8> No cure for the common cold
    exists , but the symptoms can be treated .
EXPLANATION: The <FKGL=4.0> token specifies that the
    target Flesch-Kincaid Grade Level should be
    approximately 4.0. Lower values indicate simpler
     text.<|eot_id|><|start_header_id|>assistant<|
    end_header_id|>

<FKGL=4.0><|eot_id|>
```

(b) `Llama-3.1-8B-Instruct`

```
<|begin_of_text|><|start_header_id|>system<|
    end_header_id|>

Cutting Knowledge Date: December 2023
Today Date: 19 May 2025

You are a helpful assistant. You are an expert in
    controlled text simplification. When you receive
     a text, you simplify it by rewriting it in a
    manner that is easier to read. Your
    simplification is guided by the simplification
    criteria specified by the user. You generate
    only the simplification result, without any
    additional comments or explanations.<|eot_id|><|
    start_header_id|>user<|end_header_id|>

INSTRUCTION: Simplify the following text such that
    its Flesch-Kincaid Grade Level (FKGL) score is
    approximately equal to that specified in the
    control token prepended to your generated
    simplification. The control token has the
    following format: <METRIC=VALUE>. The control
    token prepended to the source text indicates the
     FKGL value of the source text.
SOURCE TEXT: <FKGL=4.8> No cure for the common cold
    exists , but the symptoms can be treated .
EXPLANATION: The <FKGL=4.0> token specifies that the
    target Flesch-Kincaid Grade Level should be
    approximately 4.0. Lower values indicate simpler
     text.<|eot_id|><|start_header_id|>assistant<|
    end_header_id|>

<FKGL=4.0><|eot_id|>
```

(c) `Mistral-3B-Instruct`

```
<s>system
You are a helpful expert in text simplification. You
    generate a simplified version of the text input
    by the user. You simplify the text according to
    the instructions given by the user. When asked
    to simplify a text, generate only the requested
    simplification, without any additional comments,
     notes or explanations.</s>
<s>user
INSTRUCTION: Simplify the following text such that
    its Flesch-Kincaid Grade Level (FKGL) score is
    approximately equal to that specified in the
    control token prepended to your generated
    simplification. The control token has the
    following format: <METRIC=VALUE>. The control
    token prepended to the source text indicates the
     FKGL value of the source text.
SOURCE TEXT: <FKGL=4.8> No cure for the common cold
    exists , but the symptoms can be treated .
EXPLANATION: The <FKGL=4.0> token specifies that the
    target Flesch-Kincaid Grade Level should be
    approximately 4.0. Lower values indicate simpler
     text.</s>
<s>assistant
<FKGL=4.0> </s>
```

(d) `Mistral-7B-Instruct-v0.1`

```
<s> [INST] You are a helpful expert in text
    simplification. You generate a simplified
    version of the text input by the user. You
    simplify the text according to the instructions
    given by the user. When asked to simplify a text
    , generate only the requested simplification,
    without any additional comments, notes or
    explanations.

INSTRUCTION: Simplify the following text such that
    its Flesch-Kincaid Grade Level (FKGL) score is
    approximately equal to that specified in the
    control token prepended to your generated
    simplification. The control token has the
    following format: <METRIC=VALUE>. The control
    token prepended to the source text indicates the
     FKGL value of the source text.
SOURCE TEXT: <FKGL=7.2> The common distance of the
    points of a circle from its center is called its
     radius .

EXPLANATION: The <FKGL=3.6> token specifies that the
    target Flesch-Kincaid Grade Level should be
    approximately 3.6. Lower values indicate simpler
     text. [/INST] <FKGL=3.6> </s>
```

(e) `Qwen3-1.7B` & `Qwen3-8B`

```
<|im_start|>system
You are a helpful assistant. You are an expert in controlled text simplification. When you
    receive a text, you simplify it by rewriting it in a manner that is easier to read. Your
    simplification is guided by the simplification criteria specified by the user. You generate
    only the simplification result, without any additional comments or explanations.<|im_end|>
<|im_start|>user
INSTRUCTION: Simplify the following text such that its Dale-Chall Readability Score is
    approximately equal to that specified in the control token prepended to your generated
    simplification. The control token has the following format: <METRIC=VALUE>. The control
    token prepended to the source text indicates the Dale-Chall value of the source text.
SOURCE TEXT: <DALE-CHALL=8.21> The boundaries have been drawn to take in as much as possible of
    the course of the brook and linking five distinct zones.
EXPLANATION: The <DALE-CHALL=8.21> token specifies that the target Dale-Chall readability score
    should be approximately 8.21. Lower values indicate simpler text.<|im_end|>
<|im_start|>assistant
<think>

</think>

<DALE-CHALL=8.21> <|im_end|>
```

F. Training Configuration

Fine-tuning and inference experiments were conducted on the University of Zurich Science Cluster, using NVIDIA A100 GPUs with 80 GB memory. Each model was trained and evaluated on a single GPU, without distributed training. This setup was consistent across all model families and size to ensure comparable experimental conditions for both full and LoRA-based fine-tuning. A subset of inference runs was additionally conducted on the UBELIX cluster (University of Bern), using the same GPU configuration.

Table 7: Left: training hyperparameters selected for smaller ($\leq$4B) models, fine-tuned without PEFT. Right: training hyperparameters selected for larger ($>$4B) models, fine-tuned with PEFT.

Hyperparameters	**Shared Hyperparams**	
batch size	4	
grad. accumulation steps	4	
cumulative batch size	16	
weight decay	0.01	
warmup steps	30	
max epochs	3	
scheduler	cosine	
optimizer	AdamW	
max length	512 (Newsela: 4096)	
Hyperparameters	**No PEFT**	**PEFT**
learning rate	5e-6	1e-4
max grad norm	0.5	1.0
patience	4	3
LoRA rank	–	8
LoRA alpha	–	16
LoRA dropout	–	0.1

Table 8: Random seeds grouped by experiment, with results are aggregated across multiple.

Experiment	**Seeds**
Strat. Partitioning	2746317213, 478163327, 107420369, 3184935163, 1181241943, 1051802512, 958682846, 599310825, 3163119785, 440213415
Downsampling	69, 1, 40, 7, 29, 48, 78, 34, 67, 84
Robust LLM Eval.	37, 15, 96, 2, 28

Table 9: Hyperparameter optimization on Weights&Biases platform. Left: sweep configuration for smaller ($\leq$4B) models, finetuned without PEFT. Right: sweep configuration for larger ($>$4B) models, finetuned with PEFT (LoRA).

No PEFT

```
method: grid
metric:
  goal: minimize
  name: eval_loss
parameters:
  learning_rate:
    values:
      - 1e-06
      - 5e-06
      - 1e-05
      - 5e-05
  max_grad_norm:
    values:
      - 0.5
      - 1
      - 2
  weight_decay:
    values:
      - 0
      - 0.01
      - 0.1
```

PEFT

```
method: grid
metric:
  goal: minimize
  name: eval_loss
parameters:
  learning_rate:
    values:
      - 1e-05
      - 5e-05
      - 0.0001
      - 0.0002
  lora_dropout:
    values:
      - 0.05
      - 0.1
  lora_r:
    values:
      - 4
      - 8
      - 16
  max_grad_norm:
    values:
      - 0.5
      - 1
```

PLABA-EVAL: A Multi-Dimensional, In-Context Sentence Readability Dataset for Medical Text

Kexin Bian[1], **Su-Youn Yoon**[1,2], **Mamoru Komachi**[1]
[1]Hitotsubashi University [2]EduLab, Inc.
{kexin, komachi}@scl.sds.hit-u.ac.jp, su-youn.yoon@edulab-inc.com

Abstract

We introduce PLABA-EVAL, a dataset for in-context sentence-level readability assessment in biomedical abstracts and their plain-language adaptations. Participants read biomedical abstracts with full-document access, provide sentence-level ratings of *Processing Ease* and *Perceived Understanding*, and then complete an open-book multiple-choice comprehension check. The dataset comprises 609 sentences from 78 biomedical abstracts and expert plain-language adaptations, each read and rated in the context of its full document by three independent raters, together with responses to 168 manually written open-book MCQs. Our work complements existing medical simplification and readability resources that focus on lexical simplification, single-score readability labels, or decontextualized sentence ratings. Analyses show that ease and understanding can diverge at the sentence level, that simplification yields uneven gains across sentences, and that high perceived understanding does not eliminate comprehension errors. We further provide baseline linguistic analyses and comparisons to existing readability predictors, illustrating how PLABA-EVAL can support work on readability assessment, simplification, and sentence-level difficulty modeling.

Keywords: readability assessment, text simplification, human evaluation

1. Introduction

Biomedical research is increasingly consumed by non-specialists, including patients, caregivers, and the general public (Attal et al., 2023). In these settings, accessibility is not only about scientific accuracy but also about whether readers can extract key claims efficiently and understand them correctly. Readability modeling and text simplification aim to support this goal, yet many practical tools and NLP benchmarks treat difficulty as a single document-level score (Vajjala and Meurers, 2012; Xu et al., 2015; Vajjala and Lučić, 2018). This is a poor fit for how people read, as difficulty is often experienced locally, where a mostly accessible abstract can still contain a handful of sentences that are disproportionately effortful or confusing. For sentence-level methods, difficulty is often estimated from sentences presented in isolation. This setting removes discourse context that readers rely on for interpretation, and can therefore misrepresent the difficulty a sentence poses as part of the full text (Schumacher et al., 2016; Iavarone et al., 2021).

A second gap is that a single "difficulty" judgment conflates distinct aspects of reader experience. Sentences can be easy to process yet remain underspecified or confusing in context, and conversely can require effort but still be understood after integration (Van den Broek et al., 1999). Moreover, perceived understanding can diverge from their actual comprehension, motivating the use of explicit comprehension checks (Cohen et al., 2025; Leroy et al., 2012). Existing resources rarely capture these distinctions, making it difficult to disentangle ease from perceived and actual understanding, and to characterize how simplification affects each dimension across sentences.

To address these gaps, we introduce an in-context, multi-dimensional approach to sentence-level readability assessment for medical text, while combining local ratings with an open-book comprehension check. Our contributions are as follows:

- We introduce **PLABA-EVAL**, a dataset for in-context sentence-level readability assessment in biomedical abstracts and expert plain-language adaptations, with full-document sentence ratings and manually curated MCQs that assess comprehension.[1]
- We present a multi-dimensional view of readability that distinguishes *Processing Ease*, *Perceived Understanding*, and actual comprehension as measured by MCQs.
- We provide analyses and baselines showing that these dimensions are related but not interchangeable, that simplification yields uneven sentence-level gains, and that existing readability predictors capture only part of the human signal.

2. Related Work

Readability is traditionally defined as the ease with which text is read and understood (Dale and

[1]We release PLABA-EVAL under CC BY 4.0 to enable broad reuse and redistribution with attribution.

Proceedings of the Joint Workshop on Readability and Text Simplification (READIxTSAR) @ LREC 2026, pages 49–60
11 May 2026.

Chall, 1948; Richards and Schmidt, 2013). Cognitive and reading theories suggest, however, that these are not identical: online processing effort and successful meaning construction can diverge (Van den Broek et al., 1999; Just and Carpenter, 1980; van den Broek et al., 2011). Subjective perceptions of difficulty may likewise differ from comprehension performance (Leroy et al., 2010), especially when readers form only shallow or incomplete interpretations (McKoon and Ratcliff, 1992; Trabasso and Van Den Broek, 1985). In health contexts, simplification has been shown to lower perceived difficulty without consistently improving objective information retention (Leroy et al., 2010; Shulman et al., 2020). These distinctions motivate combining subjective ratings with explicit comprehension checks.

These measurement considerations also interact with the choice of annotation unit. Prior work has motivated sentence-level complexity or understandability ratings, but discourse-level factors such as cohesion can shape comprehension across sentences and paragraphs (Snow, 2002). Fully decontextualized sentence judgments may therefore be incomplete, especially for lay readers of technical medical materials, where successful integration depends on domain knowledge and how explicitly relations are stated (Ozuru et al., 2009; Kindig et al., 2004; Berkman et al., 2011). This motivates sentence-level annotation with access to surrounding context.

Existing resources relevant to sentence difficulty emphasize different aspects of the problem. Among those that incorporate human feedback, many collapse complexity into a single scale, such as CEFR-aligned rankings in MedReadMe (Jiang and Xu, 2024) or grade-level targets in general-domain datasets like CLEAR (Crossley et al., 2023). However, CEFR in particular was designed for language-learner proficiency, making it a less natural fit for lay medical comprehension (Council of Europe, 2001; Crossley et al., 2023).

Many resources focus on identifying difficult terms and how they should be simplified (Ondov et al., 2026; Xia et al., 2025), which is highly relevant in medical settings, where lexical accessibility is a major source of difficulty. But lexical difficulty alone does not fully determine sentence difficulty in context. Sentence-difficulty resources that manipulate local contextual windows (Schumacher et al., 2016; Iavarone et al., 2021) show that surrounding context can change perceived difficulty, but they typically model context through limited local windows rather than full-document reading context. Finally, a smaller line of work on health-information readability and simplification has also incorporated comprehension questions to distinguish perceived from actual difficulty (Leroy et al., 2010; Guidroz et al., 2025). However, these studies do not typically pair comprehension checks with in-context sentence-level ratings. Our work aims to bridge these gaps and provide a more granular view of how readers process and understand medical text.

	Original		Simplified	
	Mean	SD	Mean	SD
Sent per doc	6.97	1.81	8.64	2.96
Word per doc	152.36	14.47	174.49	38.31
MedReadMe	4.67	0.40	4.44	0.39

Table 1: Descriptive statistics for original and simplified documents.

3. Data Collection

We construct PLABA-EVAL by sampling 39 PubMed abstracts and their corresponding expert adaptations from the PLABA corpus (Attal et al., 2023). This results in a parallel set of 78 documents, totaling 609 sentences (272 in the original abstracts and 337 in the adaptations). We collect 3,654 sentence-level Ease/Understanding ratings from three independent raters per document, along with 468 document-level ratings. To evaluate comprehension, we additionally develop 168 open-book MCQs answerable from both variants.

3.1. Source stimuli

We draw our source stimuli from the Plain Language Adaptation of Biomedical Abstracts (PLABA) dataset (Attal et al., 2023), which pairs PubMed abstracts with manual plain-language adaptations written to answer high-frequency consumer health questions from MedlinePlus query logs.

We applied filters to retain abstracts between 120 and 180 words and applied stratified sampling over document-level readability, computed as the mean MedReadMe sentence score per abstract (Jiang and Xu, 2024) [2], to cover a range of baseline difficulty. To encourage topical diversity, we limited the sample to at most two abstracts per PLABA source question, yielding 39 source–adaptation pairs. Table 1 reports summary statistics for the sampled subset.

3.2. Text Difficulty Rating Procedure

Figure 1 summarizes the procedure. Participants were instructed to read each abstract naturally, as if

[2]We use the authors' released checkpoint: https://huggingface.co/chaojiang06/medreadme_medical_sentence_readability_prediction_CWI

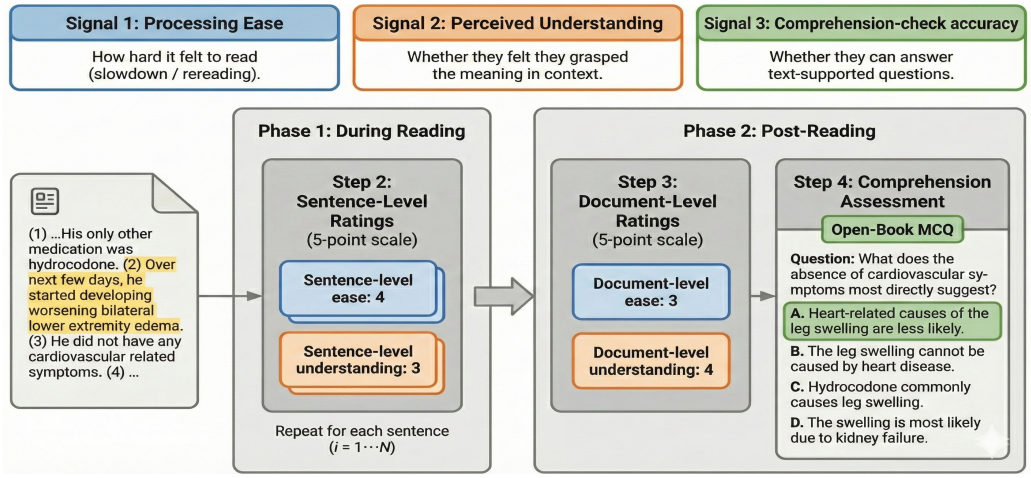


Figure 1: Difficulty rating framework. Participants read each document with full-text access and provide sentence-level ratings in sequence on two 5-point scales. After reading, they provide document-level ratings on the same scales and complete an open-book multiple-choice comprehension check.

looking for important medical information for themselves or a family member. They then completed a short onboarding tutorial to familiarize them with the interface, rating scales, and study flow.

We elicited text-difficulty judgments at two points: during reading and post-reading, followed by an open-book comprehension check. During reading, participants rated each sentence in sequence on two 1–5 scales and provided brief diagnostic feedback, with the full document visible at all times. After completing all sentence-level ratings for a document, participants provided corresponding document-level ratings. To characterize potential variation in reader background, we also collected self-reported topic familiarity and familiarity with academic-style writing. Finally, participants answered text-supported multiple-choice questions (MCQs) with the document still visible; correct answers were revealed after submission to avoid leaving participants misinformed by any study item.

3.3. Measures

Subjective ratings We operationalize perceived difficulty using two complementary subjective signals collected on 1–5 scales (higher is better): *Processing ease* ("While reading, how easy was this sentence to read?"), capturing experienced workload (e.g., slowdown or rereading); *Perceived understanding* ("How confident are you that you understood what this sentence means in this context?"), capturing metacognitive confidence that the sentence "made sense" given the surrounding text.

Diagnostic feedback After rating each sentence, participants also provided diagnostic feedback on why a sentence felt difficult. In the first half of data collection, raters could optionally leave brief free-text comments (e.g., to flag confusing terms or ask questions). In the second half, raters selected up to three difficulty factors from a fixed set of *issue tags* derived from earlier free-text feedback (e.g., vocabulary, structure/density, unnatural wording, unclear takeaway, unclear logic/connection).

Comprehension check (MCQs) After rating each document, participants answered an open-book comprehension check consisting of multiple-choice questions designed to test text-supported understanding. Items targeted information likely to matter to lay readers and avoided questions that hinge solely on highly technical details. The open-book format reduces memory demands and shifts the task toward locating and using textual evidence, so errors more directly reflect misinterpretation or failure to apply relevant evidence (Durning et al., 2016).

3.4. MCQ Curation

We manually wrote MCQs and used the same question set for both versions within each original–simplified abstract pair. To support this paired design, items targeted claims preserved under simplification, and we aimed to minimize lexical overlap between item text and either abstract version to avoid version-specific cues.

Because the task was open-book, stems and correct answers were often paraphrased to discourage keyword matching. Distractors were designed as plausible alternatives, drawing on common sources of confusion such as superficial lexical similarity, reasonable but incorrect interpretations, or scope/magnitude changes. All items were manually drafted and reviewed to ensure answerability from the abstract and exactly one correct option.

LLMs were used as a brainstorming aid for a subset of items (e.g., candidate stems or distractors); all suggestions were treated as drafts and manually edited and verified against the text.

Quality check We further validated MCQ quality using GPT-5.2, prompting the model to answer using only the provided abstract or simplification. The model agreed with the manually verified gold answers on 314/316 item–version instances (99.37%). The two discrepant cases, both on simplified abstracts, were traced to subtle interpretive differences in paraphrasing. After manual review, these items were excluded from the final dataset and all subsequent analyses.

Critical sentence Following evidence-linking practices in MCQ reading comprehension (e.g., STARC (Berzak et al., 2020)), each item was anchored to a single critical *evidence sentence*. When evidence was distributed or required cross-sentence integration, we selected the sentence that most explicitly supported the correct option. We recorded sentence indices in the original abstracts and mapped them to the simplified versions using PLABA sentence alignments.

3.5. Participants and Setup

To approximate reading by non-expert consumers of health information, we recruited native English speakers without screening for biomedical background. To characterize potential variation in expertise, we collected self-reported topic familiarity (per abstract) and familiarity with academic-style writing (frequency of reading academic texts). Because topic familiarity depends on document topic, we treat it as descriptive and not directly comparable across topics. Overall, a majority of raters reported that most or almost all information in the text was new to them, consistent with a largely non-expert reader pool. Some raters reported relevant personal experience (e.g., being diagnosed with the condition discussed); we retain these cases as plausible in real-world health-information seeking.

Implementation Participants were recruited via Prolific with eligibility restricted to first-language English speakers residing in US/UK/CA/AU and approval rate > 0.99, but did not otherwise control for socio-professional background. The study was implemented as a web-based annotation interface. Each document was annotated by three distinct workers, recruited independently for each task instance (i.e., no worker overlap across documents). Compensation was £1.50 for an estimated completion time of 8 minutes; the median observed completion time was 10.50 minutes (IQR: 6.26).

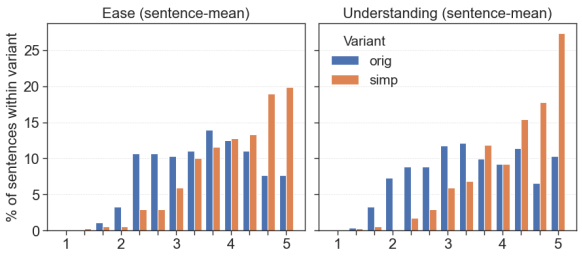


Figure 2: Sentence-level distributions of subjective ratings by variant (mean of three rater judgments). Bars show the percentage of sentences in each bin for the original (n=272) and simplified (n=337) variants.

3.6. Data Quality

Because the dataset intentionally captures subjective reading experience, we incorporated design-time safeguards to improve comparability across raters, texts, and variants. Instructions framed the task as goal-oriented health-information reading (Sabou et al., 2014), and the 1–5 rubrics for ease and understanding were anchored with behavioral examples (Melchers et al., 2011). Workers completed a short practice round before the main task (Knowles and Lo, 2025). We excluded low-quality submissions using completion-time thresholds and response-pattern checks (e.g., straightlining).

Reliability / Agreement We estimate inter-rater reliability using Gwet's AC2 with quadratic weights, which is comparatively robust to prevalence effects (Gwet, 2014). Computed per document and summarized across documents, median AC2 is 0.53 for Ease and 0.52 for Understanding, with substantial between-text variability (Ease IQR $= 0.40$, Understanding IQR $= 0.38$). This level of agreement is consistent with a subjective judgment task and with heterogeneity across biomedical texts. We find no evidence that agreement differs between original and simplified variants ($p = 0.145$), suggesting that the protocol yields comparable reliability across conditions.

4. Behavioral Findings

4.1. Subjective Rating Distributions

Sentence-level rating distributions Figure 2 shows sentence-level ratings by variant (mean over three raters per sentence). For the original abstracts, ratings span most of the 1–5 scale but are concentrated in the mid-to-high range, with relatively few sentences receiving very low scores on either dimension. Simplification shifts the distributions rightward on both axes (Ease $3.53 \pm 0.87 \rightarrow 4.08 \pm 0.79$; Understanding $3.46 \pm 0.97 \rightarrow 4.21 \pm 0.77$,

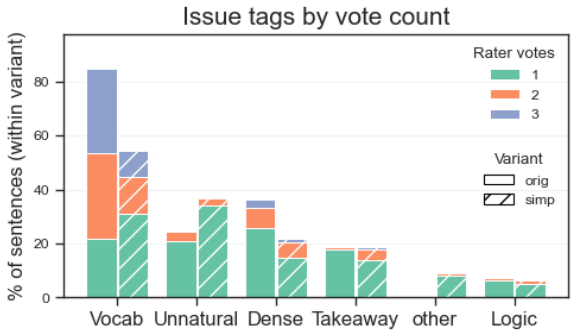


Figure 3: Issue tag prevalence by variant, shown as the percentage of sentences (within each variant) with 1, 2, or 3 raters endorsing each tag.

orig→simp), and increases near-ceiling ratings, most notably for Understanding (with a clear mass at $= 5$).

Issue tag prevalence Figure 3 summarizes issue-tag prevalence by variant, stratified by the number of raters endorsing each tag. Vocabulary-related issues are the most frequent in both variants, with substantially stronger multi-rater agreement in the original text (a larger 2–3 vote component) than in the simplified text. Sentence density is flagged more often in the original variant, whereas Unnaturalness/Redundancy is flagged more often in the simplified variant. For most non-vocabulary tags (e.g., LOGIC and OTHER), endorsements are dominated by single-rater votes, indicating weaker consensus.

Document-level rating distributions Overall, post-reading judgments on documents are more moderate. At the document level, mean ratings are less concentrated at the higher end than at the sentence level, and occupy a broader range of values. By variant, document-level averages are also higher in the simplified condition (Ease $2.90 \pm 1.04 \rightarrow 3.47 \pm 1.12$; Understanding $3.15 \pm 1.09 \rightarrow 3.89 \pm 1.06$, orig→simp), but the ceiling effect is substantially weaker than for sentence-level Understanding.

We also quantify within-abstract variation by summarizing the spread of sentence-level mean ratings across sentences (P90–P10) within each document. The resulting ranges are typically 1.2–1.5 points on a 5-point scale, indicating that most abstracts contain a mix of locally easier and more challenging sentences rather than a uniform difficulty level. Document-level Ease and Understanding align strongly with the mean sentence rating (Spearman $\rho \approx 0.77$–0.83), suggesting that global ratings reflect an integrated impression across sentences.

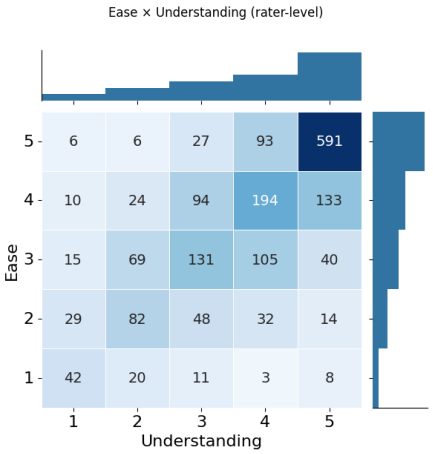


Figure 4: Joint distribution of rater-level Ease and Understanding scores (pooled across variants). Cells show counts of individual judgments; marginal bars show the corresponding one-dimensional distributions, highlighting ceiling effects.

4.2. Relationship between Ease and Understanding

Figure 4 shows the joint distribution of rater-level Ease and Understanding judgments. Ratings concentrate along the diagonal, with a strong overall association (rater-level Spearman $\rho = 0.696$; sentence-mean $\rho = 0.772$), indicating that sentences judged easier are typically also judged better understood. The positive association also holds within each variant (orig $\rho = 0.684$; simp $\rho = 0.656$).

However, the measures are not interchangeable. A nontrivial share of judgments falls off the diagonal, with large discrepancies persisting (orig: 12.3%, simp: 9.5% differ by ≥ 2 points in absolute value). Qualitatively, divergences often reflect a mismatch between decoding load and inferential load. For instance, brief clinical statements (e.g., “She was intubated for airway protection.”) are rated easy to read but poorly understood without domain knowledge, whereas sentences that add explanatory scaffolding (e.g., “hyperferritinemia (high levels of ferritin – an iron-containing protein)”) are rated harder to read yet better understood. These divergences reflect distinct sources of difficulty, which call for different interventions, motivating multi-signal evaluation rather than a single readability score.

4.3. Effects of Simplification on Sentence Ratings and Issue Tags

Ratings We estimate variant differences using between-condition comparisons on sentence alignments between the original and simplified variants ($n = 265$). Alignments include one-to-many cases arising from sentence splitting; in these cases, the

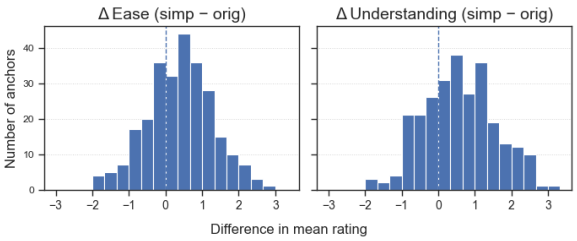


Figure 5: Alignment-based rating changes under simplification. Histograms show Δ = simp − orig in sentence-mean Ease and Understanding across sentence alignments between the original and simplified variants ($n = 265$)

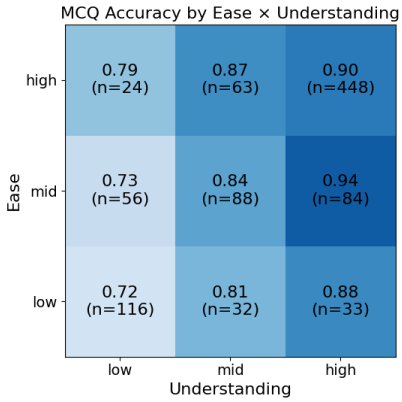


Figure 6: MCQ accuracy by evidence-sentence ratings. Cells show $P(\text{correct})$ (and n) by binned Understanding × Ease (low=1–2, mid=3, high=4–5).

simplified score is averaged across the corresponding split sentences.

Figure 5 shows the distribution of alignment-based changes in sentence-mean ratings (Δ = simp − orig). Both Ease and Understanding exhibit a clear positive shift, with median $\Delta = 2/3$ (mean $= 0.54$ for Ease and 0.72 for Understanding), indicating larger gains for Understanding. Large joint improvements are frequent (100/265; 37.7%), where both dimensions improve by at least $2/3$ ($\Delta_{\text{Ease}} \geq 2/3$ and $\Delta_{\text{Und}} \geq 2/3$). Among one-dimensional improvements (one dimension $\geq 2/3$ while the other below threshold), shifts are more common for Understanding than for Ease (14.0% vs. 6.4%). Notably, simplification is not uniformly beneficial: 27.2% show no meaningful change on either dimension, and 14.7% worsen on at least one dimension. This heterogeneity suggests caution in treating human-written simplifications as uniformly improved gold references for sentence-level evaluation.

Issue tags To characterize *what* changes when ratings shift, we analyze how issue-tag endorsements change under simplification for sentence alignments from the second half of data collection ($n = 123$), when raters selected up to three tags from a fixed set. For each tag, we compare the number of raters endorsing it in the original and simplified versions of the same aligned sentence and summarize the change as Δvotes = simp − orig.

Overall, simplification most consistently reduces terminology-related difficulty. UNCLEAR TERMS shows a large negative shift in endorsement (mean Δvotes $= -0.98$, computed over the 109 aligned pairs where the tag appears in either variant). SENTENCE DENSITY also tends to diminish, but less uniformly (mean Δvotes $= -0.40$; $n = 50$). In contrast, simplification more often introduces phrasing-level costs: UNNATURAL/WORDY PHRASING shifts upward (mean Δvotes $= +0.18$; $n = 61$). TAKEAWAY CLARITY is comparatively unstable, showing near-symmetric increases and decreases (mean Δvotes $= +0.14$; $n = 43$). These shifts are illustrated by paired examples in Table 2, highlighting how simplification can remove shared lexical barriers while leaving (or occasionally introducing) more heterogeneous clarity or coherence concerns.

4.4. Objective Comprehension (MCQ)

Across all MCQ responses (N=942), accuracy is high at 0.857. This ceiling-adjacent performance limits sensitivity to small comprehension differences, since multiple-choice correctness can reflect partial knowledge, elimination strategies, or guessing. However, MCQ accuracy still provides an objective check on self-reports and supports within-item comparisons across variants.

Relationship between subjective ratings and MCQ accuracy Figure 6 relates MCQ accuracy to workers' Ease and Understanding ratings for the corresponding evidence sentence. Accuracy increases monotonically with Understanding across Ease bins, supporting perceived understanding as a meaningful subjective proxy for text-supported comprehension. While the heatmap suggests a small advantage for higher Ease, we find no clear additional benefit of Ease beyond Understanding.

Echoing prior findings that subjective difficulty is not equivalent to comprehension performance (Leroy et al., 2013), we find that even when raters report high perceived understanding, MCQ errors persist. We verified that these errors are not driven by a single unusually difficult question or a small subset of workers. Qualitative inspection suggests that many such errors occur on implication-style items rather than fact-retrieval questions. One possible explanation is that perceived understanding and MCQ accuracy diverge when readers form a good-enough local interpretation without fully encoding the precise relation, scope, or implication re-

Original	Simplified	Issue Tag votes
Mutations in an enzyme, such as PAH, are recessive since one functioning enzyme with the wild-type allele is sufficient.	Changes to DNA of a protein, such as PAH, are made in both copies of the gene that is altered, because one working copy of the gene allows the protein to function.	Vocab 1 (-2); Dense 1 (± 0); Unnatural 1 ($+1$); Takeaway 2 ($+2$);
Once the AF frequency has been estimated and tracked by a hidden Markov model approach, the resulting trend is analyzed for the purpose of detecting and characterizing the presence of circadian variation.	When the atrial fibrillation frequency is estimated and tracked by signal processing tools, the information is further reviewed to detect and describe the presence of circadian variation.	Vocab 1 (-2); Dense 1 (-2); Unnatural 1 (± 0); Reference 1 ($+1$)

Table 2: Aligned sentence pairs illustrating how simplification changes *what readers flag as difficult*. The right column reports issue-tag votes (out of three raters) for the simplified sentence, with change $\Delta = \text{simp} - \text{orig}$; negative values indicate fewer raters endorsing the issue after simplification.

quired by the question (McKoon and Ratcliff, 1992). More targeted follow-up designs, such as collecting self-explanations or think-aloud reports, could help better characterize when subjective understanding aligns with, or diverges from, successful comprehension.

Effect of simplification on MCQ accuracy We estimate simplification effects on comprehension via within-item comparisons of MCQ accuracy, using the same question set for both variants of each base text. Overall, accuracy is modestly higher in the simplified condition (orig: 0.84 vs. simp: 0.88). At the item level, we pair 154 questions that appear in both variants and compare per-item accuracy (mean over three respondents per version). 89 items (58%) show no change in accuracy, partly due to ceiling effects. Among the remainder, accuracy improves for 40 items; 9 show large gains (i.e., $\geq$2 additional correct responses out of 3). Many of these large-gain items involve correctly identifying specific terms or relations stated in the text, suggesting that simplification helps most when comprehension is limited by lexical/terminological load. 24 items show a drop in accuracy under simplification. In several cases, the simplified wording appears to broaden or shift key terms, making the evidence sentence a less direct cue for the intended inference and leaving more room for distractors, even though the question remains answerable.

For each question, we also collected two 1–5 self-reports of perceived ease of answering and confidence in the selected answer. Beyond correctness, participants report higher ease and confidence when answering the MCQs after reading simplified texts (Δquiz-ease $\approx +0.19$; Δconfidence $\approx +0.42$, simp–orig), consistent with prior work that simplification also improves subjective experience during comprehension tasks (Guidroz et al., 2025).

5. Correlates and Predictors of Difficulty

We examine how readers' sentence-level ratings of Ease and Perceived Understanding relate to both linguistic features and existing readability predictors. For the feature-based analyses, we extract variables from established toolkits (e.g., LFTK (Lee and Lee, 2023), LCA (Lu, 2012), SCA (Lu, 2010)) as well as coherence- and cohesion-oriented measures linking each sentence to its context. We exclude features with low coverage or negligible variation and use FDR correction within each outcome (Benjamini and Hochberg, 1995). For each linguistic feature, we run a separate regression predicting the sentence's mean rating, including variant and document fixed effects; features are z-scored and standard errors clustered by document.

5.1. Sentence-level Linguistic Features

Table 3 summarizes representative correlates of sentence-level Ease and Perceived Understanding after redundancy pruning. Overall, our linguistic correlates are broadly consistent with previous findings in that **lexical accessibility** emerges as the strongest signal across both outcomes (Xia et al., 2025; Jiang and Xu, 2024). In particular, *jargon count* is among the clearest correlates of lower Ease and Perceived Understanding, suggesting that biomedical terminology is a major driver of reader-perceived difficulty in our data. Several top predictors reflect local lexical bottlenecks, such as low *minimum word frequency* or high *maximum Age of Acquisition*, indicating that sentence difficulty in our data is sensitive not only to aggregate complexity but also to especially demanding lexical items.

Beyond vocabulary, features related to **information packaging** and **surface form**, such as *nominal density* and *long-distance syntactic dependencies*, are also associated with lower ratings. This is consistent with greater difficulty when in-

Table 3: Representative Linguistic Features with Effect Sizes. Coefficients are per one-standard-deviation increase in the feature, with document and variant fixed effects and document-clustered standard errors.

Feature	β_{ease}	$\beta_{und.}$
Num. of jargons[a]	-0.437	-0.372
Words $\geq$3 syllables	-0.373	-0.305
Min word frequency	0.324	0.284
Avg word frequency	0.288	0.282
Lexical sophistication (LS2)	-0.275	-0.254
Max AoA	-0.288	-0.218
Max dep. link length	-0.302	-0.209
Complex nominals / T-unit	-0.267	-0.184
Mean length of clause	-0.216	-0.141
Punctuations / sentence	-0.295	-0.206

[a] Jargon spans are extracted using the released complex-span identification checkpoint from (Jiang and Xu, 2024): https://huggingface.co/chao06/medreadme_medical_complex_span_identification_CWI.

formation is packed into complex noun phrases or spread across long spans (Pitler and Nenkova, 2008). *Punctuation density* shows a similar pattern, plausibly reflecting parentheticals or stacked clauses that fragment the sentence and increase reading difficulty.

Despite substantial overlap in their strongest correlates, Ease and Understanding differ in how they relate to specific linguistic features. We additionally fit a stacked regression with a feature$\times$outcome interaction (Ease vs. Understanding), and treat cross-outcome differences as reliable only when the interaction survives FDR correction. Several of the clearest differences involve **reference and grounding**. *Third-person singular pronoun incidence* is positively associated with both outcomes but more strongly with Ease ($\Delta = \beta_{\text{und}} - \beta_{\text{ease}} \approx -0.09$). This is in line with prior work showing that pronouns can reduce local processing cost relative to repeated names when the referent is already discourse-salient (Gordon et al., 1993), even if resolving referents can still leave uncertainty about meaning. In contrast, higher *named-entity density* is linked to lower Understanding than Ease ($\Delta \approx -0.09$), suggesting that unfamiliar entities can reduce confidence in understanding even when decoding is not especially difficult. We also find systematic differences for **lexical variability** and **syntactic packaging**. For example, high *textual lexical diversity* (McCarthy and Jarvis, 2010) is associated with lower ratings on both outcomes, but more strongly with Ease than with Understanding ($\Delta \approx +0.09$), suggesting that variability primarily increases perceived reading effort. Similarly, longer clauses and deeper dependency structures track Ease more strongly than Understanding ($\Delta \approx +0.08$), consistent with dense syntactic integration increasing processing burden without proportionally reducing confidence in comprehension.

5.2. Context / Coherence Features

We next examine whether a sentence's relationship to its surrounding context predicts perceived Ease or Understanding. We replicate the approach of (Cohen et al., 2025), considering (i) language-model chain predictability (the predictability of a sentence from its preceding context) and (ii) embedding-based proximity, operationalized as similarity to adjacent sentences or to document centroids. We additionally include (iii) cohesion indices from TAACO (Crossley et al., 2019), capturing lexical overlap with the preceding sentence, connectives, and semantic similarity.

Using the same feature-wise regression setup as above, many context metrics show negative associations with Ease/Understanding in baseline screens, but these patterns largely reflect sentence-length confounding: proximity and overlap measures are strongly length-correlated, and coefficients attenuate sharply after adding word count. After length adjustment, no coherence/proximity effects remain robust for Understanding (all small; none survive FDR), and overlap-based TAACO effects similarly disappear. The main exception is a small but stable positive association between Ease and explicit additive/connective markers (e.g., conjunctions, addition), consistent with overt discourse cues modestly improving perceived readability. Overall, context/cohesion features add limited signal beyond basic length effects in these abstracts, underscoring the importance of length control when interpreting context-based metrics. We note that the embedding-based similarity measures were computed following Cohen et al. (2025) without additional whitening or mean-centering; such post-processing may reduce hubness effects in sentence embeddings, and could be explored in future work.

5.3. Existing Readability Predictors

We lastly examine how well sentence-averaged difficulty can be predicted using existing readability predictors. We consider four commonly used **unsupervised metrics**: FKGL (Kincaid et al., 1975), ARI (Smith and Senter, 1967), SMOG (Mc Laughlin, 1969), and RSRS (Martinc et al., 2021). We also evaluate **jargon-enhanced variants** of these metrics proposed by Jiang and Xu (2024), which incorporate a weighted count of medical jargon spans into the original formulas[3]. We further

[3]We use the pre-tuned weights (α) reported in Jiang and Xu (2024).

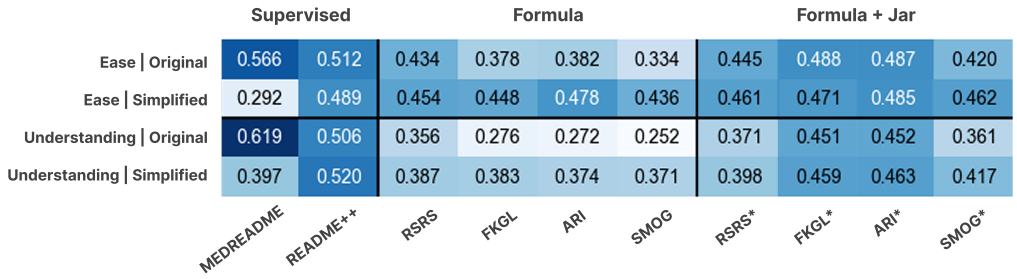


Figure 7: Alignment between readability predictors and human ratings by variant and dimension, measured with Spearman's ρ

compare against **supervised readability predictors** fine-tuned on existing readability datasets: MedReadMe (Jiang and Xu, 2024) and, as a broader-domain contrast, ReadMe++ (Naous et al., 2024). [4]

Figure 7 summarizes alignment between existing readability predictors and human ratings across both dimensions and variants. MedReadMe shows the strongest alignment on the original texts, particularly for Understanding, but its performance drops substantially on the simplified variants. We hypothesize that this reflects a transfer gap between MedReadMe's training distribution and PLABA, likely related to differences in the source and rewriting conventions underlying the simplified texts. By contrast, ReadMe++ is weaker on the originals but more stable across variants.

The impact of adding a weighted jargon-count term varies significantly by formula type. For traditional metrics (FKGL, ARI, SMOG), adding a jargon term provides a vital semantic signal that their base formulas lack, leading to substantial gains in alignment. Conversely, the neural RSRS metric see only marginal gains from explicit jargon modeling, likely because RSRS is based on word-likelihood, and already implicitly penalizes technical jargon as low-probability tokens, making manual jargon counts redundant.

The predictors also differ systematically across the two human dimensions. The base formula-based metrics align more strongly with Ease than with Understanding, consistent with their reliance on surface proxies such as length and word-form complexity. Adding jargon counts narrows this gap across all four formulas, with larger gains for Understanding than for Ease. By contrast, the MedReadMe predictor aligns more strongly with Understanding, whereas ReadMe++ is more balanced across Ease and Understanding. Overall, these patterns suggest that existing readability predictors capture part of the signal in PLABA-EVAL, but do not collapse Processing Ease and Perceived Understanding into a single construct.

[4]We use the authors' released checkpoints.

6. Conclusion

In this work, we introduced **PLABA-EVAL**, an in-context dataset and protocol for sentence-level readability assessment of biomedical abstracts and expert plain-language adaptations. By pairing sequential sentence ratings of *Processing Ease* and *Perceived Understanding* with an open-book MCQ check, the dataset supports a more fine-grained view of how readers process and understand medical text than single-score readability labels alone. Our analyses show that ease and understanding are strongly related but not interchangeable, and that existing readability predictors align inconsistently with these human judgments across dimensions and text variants. We hope PLABA-EVAL will support future work on medical readability and simplification, while encouraging more fine-grained, multi-dimensional evaluation of how texts are read and understood.

7. Limitations

A first limitation is the modest size of PLABA-EVAL. In addition, each document is annotated by only three raters, with no overlap across documents, which limits how precisely rater-specific tendencies can be separated from item difficulty. The dataset is also limited to a single genre and participant setting, so its generalizability to other medical domains, discourse structures, and reader populations remains to be established. In addition, while open-book MCQs provide an objective anchor, they capture only one form of comprehension and may miss more open-ended forms of understanding.

Finally, recent extensions to PLABA through the TREC shared task add fine-grained annotations for identifying difficult terms and characterizing replacement strategies (Ondov et al., 2026). Aligning PLABA-EVAL with these newer term-level resources would be a valuable next step for relating reader judgments to expert-annotated difficult terms and simplification strategies.

8. Acknowledgments

This work was supported by a commissioned research project from the National Institute of Information and Communications Technology (NICT), titled “Research and Development of Externally Controllable Modeling of Multimodal Information for Improving Machine Translation Accuracy” and by a Grant-in-Aid for Scientific Research from JSPS (KAKENHI Grant Number 25K03178).

9. Lay Summary

People often rely on medical texts to understand health information and make decisions, so it is important to measure whether and how a text is difficult. In practice, however, medical texts are often judged as simply “easy” or “hard.” But a sentence can be easy to read while still being hard to understand, and a sentence that takes effort to read can still make sense once the reader works through it. Moreover, people can feel that they understand a text while still missing or misunderstanding what it implies. These differences matter because they may call for different kinds of improvement, such as replacing jargon, breaking down dense sentences, clarifying how ideas are connected, or making important implications more explicit.

In this study, we created a new dataset, PLABA-EVAL, to capture these different aspects of reading difficulty in biomedical abstracts and their plain-language adaptations. Readers rated each sentence for both Processing Ease and Perceived Understanding, and also answered open-book multiple-choice questions to test what they actually understood. Our results show that ease, understanding, and comprehension do not always align, and that simplification does not improve them uniformly in every sentence. We hope this resource will help researchers build better ways of evaluating and improving medical texts so that they are not only easier to read, but also easier to truly understand.

10. Bibliographical References

Kush Attal, Brian Ondov, and Dina Demner-Fushman. 2023. A dataset for plain language adaptation of biomedical abstracts. *Scientific Data*, 10(1):8.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Nancy D Berkman, Stacey L Sheridan, Katrina E Donahue, David J Halpern, Anthony Viera, Karen Crotty, Audrey Holland, Michelle Brasure, Kathleen N Lohr, Elizabeth Harden, et al. 2011. Health literacy interventions and outcomes: an updated systematic review. *Evidence report/technology assessment*, (199):1–941.

Yevgeni Berzak, Jonathan Malmaud, and Roger Levy. 2020. STARC: Structured annotations for reading comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5726–5735.

Trevor Cohen, Weizhe Xu, Yue Guo, Serguei Pakhomov, and Gondy Leroy. 2025. Coherence and comprehensibility: Large language models predict lay understanding of health-related content. *Journal of biomedical informatics*, 161:104758.

Council of Europe. 2001. *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Scott Crossley, Aron Heintz, Joon Suh Choi, Jordan Batchelor, Mehrnoush Karimi, and Agnes Malatinszky. 2023. A large-scaled corpus for assessing text readability. *Behavior Research Methods*, 55(2):491–507.

Scott A Crossley, Kristopher Kyle, and Mihai Dascalu. 2019. The tool for the automatic analysis of cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior research methods*, 51(1):14–27.

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Steven J Durning, Ting Dong, Temple Ratcliffe, Lambert Schuwirth, Anthony R Artino Jr, John R Boulet, and Kevin Eva. 2016. Comparing open-book and closed-book examinations: a systematic review. *Academic medicine*, 91(4):583–599.

Peter C Gordon, Barbara J Grosz, and Laura A Gilliom. 1993. Pronouns, names, and the centering of attention in discourse. *Cognitive science*, 17(3):311–347.

Theo Guidroz, Diego Ardila, Jimmy Li, Adam Mansour, Paul Jhun, Nina Gonzalez, Xiang Ji, Mike Sanchez, Sujay Kakarmath, Mathias MJ Bellaiche, et al. 2025. LLM-based text simplification and its effect on user comprehension and cognitive load. *arXiv preprint arXiv:2505.01980*.

Kilem L Gwet. 2014. *Handbook of inter-rater reliability: The definitive guide to measuring the*

extent of agreement among raters. Advanced Analytics, LLC.

Benedetta Iavarone, Dominique Brunato, and Felice Dell'Orletta. 2021. Sentence complexity in context. In *Proceedings of the workshop on cognitive modeling and computational linguistics*, pages 186–199.

Chao Jiang and Wei Xu. 2024. MedReadMe: A systematic study for fine-grained sentence readability in medical domain. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2024, page 17293.

Marcel A Just and Patricia A Carpenter. 1980. A theory of reading: from eye fixations to comprehension. *Psychological review*, 87(4):329.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report.

David A Kindig, Allison M Panzer, and Lynn Nielsen-Bohlman. 2004. *Health Literacy: A Prescription to End Confusion*. National Academies Press.

Rebecca Knowles and Chi-kiu Lo. 2025. Calibration and context in human evaluation of machine translation. *Natural Language Processing*, 31(4):1017–1041.

Bruce W Lee and Jason Lee. 2023. LFTK: Handcrafted features in computational linguistics. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 1–19.

Gondy Leroy, James E Endicott, Obay Mouradi, David Kauchak, and Melissa L Just. 2012. Improving perceived and actual text difficulty for health information consumers using semi-automated methods. In *AMIA Annual Symposium Proceedings*, volume 2012, page 522.

Gondy Leroy, Stephen Helmreich, and James R Cowie. 2010. The influence of text characteristics on perceived and actual difficulty of health information. *International journal of medical informatics*, 79(6):438–449.

Gondy Leroy, David Kauchak, and Obay Mouradi. 2013. A user-study measuring the effects of lexical simplification and coherence enhancement on perceived and actual text difficulty. *International journal of medical informatics*, 82(8):717–730.

Xiaofei Lu. 2010. Automatic analysis of syntactic complexity in second language writing. *International journal of corpus linguistics*, 15(4):474–496.

Xiaofei Lu. 2012. The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, 96(2):190–208.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2021. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, 47(1):141–179.

G Harry Mc Laughlin. 1969. SMOG grading-a new readability formula. *Journal of reading*, 12(8):639–646.

Philip M McCarthy and Scott Jarvis. 2010. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

Gail McKoon and Roger Ratcliff. 1992. Inference during reading. *Psychological review*, 99(3):440.

Klaus G Melchers, Nadja Lienhardt, Miriam Von Aarburg, and Martin Kleinmann. 2011. Is more structure really better? a comparison of frame-of-reference training and descriptively anchored rating scales to improve interviewers'rating quality. *Personnel Psychology*, 64(1):53–87.

Tarek Naous, Michael J Ryan, Anton Lavrouk, Mohit Chandra, and Wei Xu. 2024. ReadMe++: Benchmarking multilingual language models for multi-domain readability assessment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12230–12266.

Brian Ondov, William Xia, Kush Attal, Ishita Unde, Jerry He, and Dina Demner-Fushman. 2026. Lessons from the TREC plain language adaptation of biomedical abstracts (PLABA) track. *Journal of Biomedical Informatics*, page 104983.

Yasuhiro Ozuru, Kyle Dempsey, and Danielle S McNamara. 2009. Prior knowledge, reading skill, and text cohesion in the comprehension of science texts. *Learning and instruction*, 19(3):228–242.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 186–195.

Jack C Richards and Richard W Schmidt. 2013. *Longman dictionary of language teaching and applied linguistics*. Routledge.

Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. 2014. Corpus annotation through crowdsourcing: Towards best practice guidelines. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 859–866.

Elliot Schumacher, Maxine Eskenazi, Gwen Frishkoff, and Kevyn Collins-Thompson. 2016. Predicting the relative difficulty of single sentences with and without surrounding context. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1871–1881.

Hillary C Shulman, Graham N Dixon, Olivia M Bullock, and Daniel Colón Amill. 2020. The effects of jargon on processing fluency, self-perceptions, and scientific engagement. *Journal of Language and Social Psychology*, 39(5-6):579–597.

EA Smith and RJ Senter. 1967. Automated readability index. *AMRL-TR. Aerospace Medical Research Laboratories (US)*, pages 1–14.

Catherine Snow. 2002. *Reading for understanding: Toward an R&D program in reading comprehension*. Rand Corporation.

Tom Trabasso and Paul Van Den Broek. 1985. Causal thinking and the representation of narrative events. *Journal of memory and language*, 24(5):612–630.

Sowmya Vajjala and Ivana Lučić. 2018. Onestopenglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 297–304.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the seventh workshop on building educational applications using NLP*, pages 163–173.

Paul van den Broek, Catherine M Bohn-Gettler, Panayiota Kendeou, Sarah Carlson, and Mary Jane White. 2011. When a reader meets a text. In *Text Relevance and Learning from Text*, pages 123–139. Emerald Publishing Limited.

Paul Van den Broek, Michael Young, Yuhtsuen Tzeng, Tracy Linderholm, et al. 1999. The landscape model of reading: Inferences and the online construction of a memory representation. *The construction of mental representations during reading*, pages 71–98.

William Xia, Ishita Unde, Brian David Ondov, and Dina Demner-Fushman. 2025. Jebs: A fine-grained biomedical lexical simplification task. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17654–17666.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Automatic Extraction of Textual and Phonemic Complexity for French Cued Speech

Magali Norré,[1,3] Brigitte Bigi,[1] Núria Gala,[1]
Ludivine Javourey-Drevet,[2] Thomas François[3]
[1] Aix Marseille Univ, LPL, CNRS (UMR 7309), Aix-en-Provence, France
[2] Univ Lille, SCALab, CNRS (UMR 9193), Lille, France
[3] Université catholique de Louvain, CENTAL, ILC, Louvain-la-Neuve, Belgium
{magali.norre, nuria.gala}@univ-amu.fr, brigitte.bigi@cnrs.fr,
ludivine.javourey@univ-lille.fr, thomas.francois@uclouvain.be

Abstract

This article presents the results of an analysis of a written corpus with the view of automatically generating it in French Cued Speech (CS). CS is a communication system developed for people with hearing impairment to complement speech reading at the phonetic level using hands. This visual communication mode uses handshapes in different positions near the face in combination with the mouthshape (called 'cues' or 'keys') to make the phonemes of spoken language look different from each other. Despite many studies demonstrating its benefits, there are few resources available for learning and practicing it, especially in French. As part of a wider project aimed at creating an online learning platform with automatically generated videos using an augmented reality system displaying a virtual coding, we propose to identify, extract, and analyze 41 textual and phonemic features that might be more complex to (de)code in French CS. For the automatic extraction of complexity, several tools are used: FABRA for readability, SPPAS for phonetization and CS key generation. The results show some strong correlations between readability features, few between phonemic variables, and few between the two types. An initial model is proposed for selecting texts to be recorded for learning French CS.

Keywords: Cued speech, hearing loss, automatic features extraction, readability

1. Introduction

About 5% of the world's population live with disabling hearing loss, including 34 million children (World Health Organization, 2021). Most of them (90%) have hearing parents (Jones et al., 1989): an oral language is used for everyday communication. Failure to hearing impacts learning speech and its intelligibility. It also has an impact in learning to read. Lip reading is not enough to disambiguate some sounds, such as the visemes in '***p**ain*' (bread), '***b**ain*' (bath), and '**m**ain' (hand) in French.

Cued Speech (CS) is a visual communication system designed to improve spoken language comprehension for people with hearing impairments by using handshapes, positions around the face, and mouthshapes (called 'cue' or 'key') to disambiguate phonetic information. Oral sounds can be represented with this code, originally developed for American English by Cornett (1967), and adapted to about 65 languages, including French with *Langue française Parlée Complétée* (LfPC or LPC). The term 'French CS' is used hereafter.

Although its usefulness is recognized (Leybaert and LaSasso, 2010), there are few studies on its learning and its complexity. Gala et al. (2024) were the first to mention that readability and phonemic variables could be combined to automatically estimate the complexity for French CS. They cited several readability and phonemic features but they only annotated the CS key frequency without establishing any correlations. This paper investigates these parameters to propose a first method for the automatic classification of French CS resources graded by level of learning complexity.

There are many studies on readability and textual complexity, including several on Alector, the French corpus analyzed in this paper. Javourey-Drevet et al. (2022) proposed an analysis of this corpus, but on a small part and not adapted for the target public here. Ormaechea and Tsourakis (2024) also analyzed this corpus with other features and a different research purpose: the comparison of automatic text simplification systems. Listenability, which aims to assess listening difficulty of spoken materials, is also relevant for us. Researchers in this field have generally investigated combining textual and phonemic features to explain listenability of materials for language learners (Kotani et al., 2014; Kotani and Yoshimi, 2017), including for French (Ozawa et al., 2024). Such phonemic variables also appears relevant to model CS communication.

As part of a wider project aimed at creating an online learning platform with automatically generated videos in French CS from spoken texts, we propose to identify, extract, and analyze their textual and phonemic features that might be more complex to (de)code in French CS. These features are

Proceedings of the Joint Workshop on Readability and Text Simplification (READIxTSAR) @ LREC 2026, pages 61–73
11 May 2026.

combined in a model unsupervised to classify automatically the learner texts for the future learning platform. The paper is organized as follows. Section 2 introduces French CS. Section 3 describes the project. Section 4 presents the corpus to be annotated. Section 5 defines the variables and methods used for corpus analysis. Section 6 reports the results. Section 7 concludes the paper.

2. French Cued Speech

CS is not a language and is not intended to replace natural sign languages. Rather, it complements them by supporting access to a spoken language when such access is required or preferred (e.g., alongside French Sign Language, LSF, for French). A CS key does not encode a whole word; it encodes phonological information. This makes initial learning relatively fast (some studies report about ten hours, although details are limited), but regular practice remains necessary. Once the system is mastered, it can be used to cue any utterance in the target spoken language, including proper nouns, neologisms, and foreign words. Parents and families need to learn CS to facilitate the child's language immersion and so that the child can benefit from the contribution of CS. There are several associations that promote and teach the French CS to different audiences, such as the ALPC in France,[1] and in Belgium.[2]

CS is based on phonemes. There are 3 possible key structures: Consonant (C), Vowel (V), or CV. In French CS, there are 8 handshapes representing (semi-)consonants (Figure 1) and 5 positions near the face representing vowels (Figure 2),[3] each representing several sounds because the moutshape disambiguates them. The side position is also used for single consonant. The handshape 5 is used for single vowel. There is also a neutral handshape and a neutral position, both used during silences.

In order to facilitate corpus annotation and automatic processing, each position and handshape is assigned a symbolic label, following conventions from prior work on automatic key generation for French CS (Bigi, 2023, 2025a; Lancien and Bigi, 2025). Handshapes are labeled using numbers as shown in Figure 1, and positions with lowercase letters as shown in Figure 2.

Only a few studies have investigated the development of technologies for the generation and learning CS keys. The Swiss A Capella Foundation developed the Text2LPC system,[4] a French-focused online tool that provides text-to-LPC conversion. To the best of our knowledge, no scientific publication describing its design or evaluation is currently available. Piquard-Kipffer (2016) presented a digital album that uses a 3D avatar as narrator in French CS for children with language difficulties and learning challenges between 3 and 12 years. The texts of the stories told by the talking and coding head are organized into several levels of linguistic complexity, based on lexical, morphological, syntactic, and semantic properties. The author only gives a few examples of sentences and complexity criteria without details. Sankar (2024) also worked on an automatic system for recognition and generation of French CS. To our knowledge, these systems are not available. However, Bigi (2025a) has proposed AutoCS in TextCueS, an open source automated system for CS key generation, also available online for French and American English.[5]

The corpus and resources for learning to cue

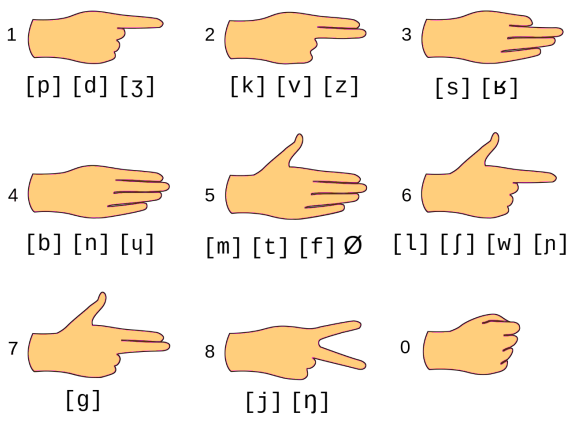


Figure 1: Handshapes representing consonants

Figure 2: Positions representing vowels

[1] https://alpc.asso.fr

[2] http://www.lpcbelgique.be

[3] For example, the fourth figure can represent the words or sounds in '*pain*' (bread), '*peu*' (few), '*deux*' (two), '*jeu*' (game), etc. because the handshape 1 is used on the cheekbone, the mouthshape makes the distinction.

[4] https://text2lpc.a-capella.ch

[5] https://auto-cuedspeech.org/textcues.html

are also scarce. Recently, Sankar (2024) built corpora in French CS, but without annotation. Another corpus is the CLeLfPC, available online on the ORTOLANG platform (Bigi et al., 2022).[6] It contains 4 hours of audio/video recordings, partly annotated. People who want to learn French CS use various resources that are not always easily available: games with speech therapists (Artaz, 2011; Olhagaray, 2013), learning CDs, short videos (e.g. some can be found on various websites and YouTube), books (Sabbagh, 2012a,b), or courses with associations.

3. The VizLector Project

The aim of the VizLector project is the creation of freely accessible resources in French CS. An online learning platform will be deployed, including video supports with humans enabling to practice and learn French CS for different audiences (e.g. children or adults with hearing impairment and their family), thus bridging the gap in resources for learners. Producing CS videos is both time-consuming and costly, as it requires the involvement of trained human cuers. Video recordings of people reading texts are processed with SPPAS (Bigi, 2015),[7] together with its AutoCS spin-off (Bigi, 2025a).[8] Using the video signal, the associated audio track, and an orthographic transcription, SPPAS automatically performs text normalization, phonetic conversion, forced alignment, and CS cue generation (keys, timing, hand positions/angles, and coordinates); the resulting hand cues are then overlaid onto the video. It can also generate the CS keys from a written text for French and American English. For written texts, the processing pipeline includes normalization, phonetization, and CS key generation (Gala et al., 2024; Bigi, 2023).

4. The Alector Corpus

For the online learning platform, it is necessary to have a written corpus with the view of automatically generating it in French CS. The 200 texts from the French Alector corpus (Gala et al., 2020)[9] were annotated and analyzed. It contains 100 original texts and their corresponding 100 adapted versions (Table 1), manually simplified by humans following the simplification guidelines developed in their project. Half of them are in the literary genre, while the other half are scientific. These texts are intended for children between 7 and 11 years (2nd to 5th grade). There are 5 text levels: IReST, CE1 (2nd), CE2 (3rd), CM1 (4th), and CM2 (5th). The first group is a set of 10 standardized, easy texts usually used for assessment of reading performances from the French version of the IReST corpus (Vital-Durand, 2011). There is a disproportion of scientific texts compared to literary texts for this level, even though there are fewer texts. We will investigate whether the level of the texts is correlated with the variables that will be extracted.

	Original		**Simplified**		**Total**
	lit	sci	lit	sci	
IReST	1	9	1	9	20
CE1	15	10	15	10	50
CE2	14	10	14	10	48
CM1	10	10	10	10	40
CM2	10	11	10	11	42
Total	50	50	50	50	200

Table 1: Number of texts per level, type and genre

5. Features Extraction

In order to develop an automatic text classification model to predict the complexity level, relevant text representations need to be extracted. In this study, two types of information were combined: readability characteristics and phonemic features. For the automatic extraction of readability features (Section 5.1), the API of FABRA (Wilkens et al., 2022) was used. FABRA is a readability toolkit that offers a large number of readability predictors for French.[10] It allows to calculate 509 language variables, most of which can be represented through 18 statistical aggregators (e.g., sum, median, average, variance, etc.). This tool can be used, for example, for corpus analysis related to readability, text simplification, automatic genre identification, etc. SPPAS v.4.29 was used for annotation of phonemic features (Section 5.2). Finally, the two types of features were combined (Section 5.3).

5.1. Readability Features

In FABRA, 30 variables were selected: 3 variables based on length (section 5.1.1), 23 lexical variables (section 5.1.2), and 5 syntactic ones (section 5.1.3). Note that 1 length-based variable not from FABRA has been added for correlations. The full list of variables used is shown in Appendix A (Table 7). It should also be mentioned that the average was used as aggregator in FABRA, except for LEXdvr-FLC, that has no underlying distribution and therefore is scalar.

[6] https://hdl.handle.net/11403/clelfpc
[7] https://sppas.org
[8] https://auto-cuedspeech.org
[9] The corpus is available on demand (https://corpusalector.huma-num.fr).
[10] https://cental.uclouvain.be/fabra

5.1.1. Length-based Variables

Length-based variables were historically the first to be used in readability to assess the difficulty of a text (Flesch, 1948; Kandel and Moles, 1958). They are still employed, despite their lack of causal relation with reading difficulty.

Word length (1 variable) and **Sentence length** (1). The number of syllables per word (LENwrdSYL), and the number of tokens per sentence (LENsntWRD) were the only length-based variables used from FABRA because the others are based on letter count, which is less relevant for CS keys (phoneme-based). More syllables and tokens require generating more CS keys, potentially resulting in more cognitive load.

Text length (1). The non-normalized length of texts (not directly output from FABRA, but reused from Alector) was added, i.e. the total number of words per text without punctuation (LENtxtWRD).

All length-based variables are used to perform correlation analyses with phonemic features.

5.1.2. Lexical Variables

Several families of lexical variables are tested.

Content overlap (1 variable). It measures repetitions of lemmas (LEXcovLGAL), we can assume that the text will be easier to (de)code because more easier to predict.[11]

Lexical diversity (1). The CTTR (Corrected Type Token-Ratio) is linked to the hapax legomena (LEXdvrFLC), the words that only occur once in a document, about 40-60% of a text (Kornai, 2007). These rare words could be more complex to (de)code. They are also longer because they are included in open word classes (nouns, verbs, etc.).

Lexical frequency (4). These variables captures the frequency of lexical words in the text, based on the lexical databases of CHILDES (LEXfrqCCS), FLELex (LEXfrqFCL), and Lexique3 (LEXfrqLCL). In addition, LEXfrqLWL considers of words in the text. More frequent words are likely to be known by a greater number of people than other words even if they do not necessarily belong to the informal language register (e.g. register used in family or classroom).

Graded lexicons (10). As with the previous variables, we can assume that texts with words assigned to the CEFR levels A1 or A2 (LEXgrd[BA1/BA2/FA1/FA2] in two lexicons) may be easier to (de)code than texts with words of level B1, B2 (LEXgrd[BB1/BB2/FB1/FB2]) or C1 and C2 (LEXgrd[FC1/FC2]). These measures are based on the Beacco's French Reference Level Descriptors and the FLELex lexicon.

Lexical norms (4). Psycholinguistic features are included, such as the age of acquisition of each word (LEXnrmAOA), word level of concreteness (LEXnrmCNCR), word familiarity or subjective frequency (LEXnrmFAM), and word imageability (LEXnrmIMG).

Lexical sophistication (3). The lexical sophistication is measured as the proportion of words in the text belonging to the first frequency bands of 1,000 words of the three following frequency lists: CHILDES (LEXsopCK1), Gougenheim vocabulary list (LEXsopGK1), and Lexique3 (LEXsopLWK1).

5.1.3. Syntactic Variables

Language development (5 variables). Since deaf students sometimes have difficulty identifying verbs in a sentence (Leitao et al., 2021), the number of words before the main verb (SYNdevBFR) is calculated. Several structural features are included: the number of constituents/phrases (SYNdevNPHRS), as well as those of the internal conjugate clause type (SYNdevNPRSSINT), of the relative clause type (SYNdevNPRSSREL), and of the subordinate clause type (SYNdevNPRSSSUB). The stories told in French CS with the avatar from the digital album by Piquard-Kipffer (2016) were also classified by level of syntactic complexity, one criterion being the greater number of complex sentences.

5.2. Phonemic Features

In total, 9 phonemic features were automatically extracted, and 1 manually annotated due to a lack of annotation tools. All phonemic variables were normalized as explained in Appendix A (Table 8).

CS key frequency (4 variables). As with Gala et al. (2024), the Alector corpus was automatically annotated with SPPAS to get the number of CS keys (PHOkey). With SPPAS, it is possible to compute the proportion of the different types of transition between face positions. We assume that transitions between more distant positions would be more complicated, such as side compared to throat and conversely (PHOpositionST) or cheekbone compared to throat and conversely (PHOpositionBT). As regards the structure of CS keys, they have three possible configurations: C, V, or CV. The proportion of CV clusters (PHOclusterCV) is automatically extracted. The CV seems to be the most complex of the three to (de)code. We assume that it is because it is necessary to consider the position of the vowel. MarsaTag (Rauzy et al., 2014) has been used to get the Part-of-Speech (POS) and observe if there are links between the most (or least) frequent tags and the number of CS keys.

[11] This was the case, for example, during the French CS training of one of the authors of the paper when it was necessary to code children's nursery rhymes containing repetitions. The participants found it easier to code.

Phonemic frequency (6). Some phonological phenomena that have an impact on the lack of phoneme-grapheme consistency – such as glides, liaisons, etc. – may seem complex to (de)code (Gala et al., 2024). Therefore, the proportion of each of the three French glides (see Table 2) was computed, i.e., the semi-consonants or semi-vowels w, ɥ, j (PHOglideW, PHOglideH, PHOglideJ). The total number of glides normalized by the number of CS keys was also considered (PHOglide). In addition, one author of the paper annotated the obligatory liaisons based on the previous study (Gala et al., 2024), their proportion was computed (PHOliaison). Liaison is a phenomenon where an orthographically-final consonant is mute except in certain environments (Table 3), i.e., when it precedes a vowel, a mute h or a glide. To our knowledge, there is no sufficiently reliable automatic tool for annotating French liaisons, in part due to the arbitrary application of liaison rules. Finally, the number of Consonant-Consonant (C-C) clusters in a same word (PHOclusterCC) was also extracted. We then assume that CS key splitting can be more complex if there is a double consonant.

Pho.	Example	Phonemes (keys)
w	***oi**seau* (bird)	wa.zo (**6**s.2s)
ɥ	*n**ui**t* (night)	n.ɥi (4s.**4**m)
j	*feu**ille*** (leaf)	fœ.j (5s.**8**s)

Table 2: Examples of glides in French

Pho.	Example	Phonemes (keys)
z	*le**s** autres* (the others)	lə.zo.t.ʁə (6t.**2s**.5s.3s)
t	*tou**t** à coup* (suddenly)	tu.ta.ku (5c.**5s**.2c)
n	*u**n** été* (one summer)	œ̃.ne.te (5t.**4t**.5t)
ʁ	*dernie**r** étage* (top floor)	dɛ.ʁ.n.je.ʁe.ta.ʒ (1c. 3s.4s.8t.**3t**.5s.1s)
p	*tro**p** agressif* (too aggressive)	t.ʁɔ.pa.g.ʁɛ.si.f (5s. 3s.**1s**.7s.3c.3m.5s)

Table 3: Examples of obligatory liaisons in French

5.3. Combining the Features

As previously mentioned, there is a lack of available training data on French CS comprehension collected from the deaf population. Therefore, unsupervised model had to be used to automatically classify texts by level of complexity. As a first step before clustering, feature selection was performed using the Minimum Redundancy Maximum Relevance (MRMR) algorithm (Ding and Peng, 2005). This method aims to identify a subset of variables that are maximally informative with respect to the target variable while minimizing redundancy among the selected features. The target variable was the text level encoded numerically to preserve its inherent ordinal nature. In v.4.3.3 of R (R Core Team, 2021), the mRMRe package (v.2.1.2.2) was used (De Jay et al., 2013), as well as the NbClust package (v.3.0.1) for determining the relevant number of clusters (Charrad et al., 2014), before applying the K-means algorithm.

6. Results

Readability features are analyzed first, then phonemic features, and finally both.

6.1. Readability Features

The number of words per text (LENtxtWRD) increases monotonically across levels of texts. Median values rose from IReST (Mdn = 127.5) to CE1 (Mdn = 244), CE2 (Mdn = 307.5), CM1 (Mdn = 370), and CM2 (Mdn = 513.5). The (average) number of syllables per word (LENwrdSYL) shows a small but consistent increase across levels. Median values ranged from 1.41 in IReST to 1.44 in CM2, with substantial overlap between distributions. Sentence length (LENsntWRD) showed moderate variation across levels. While median values increased slightly from CE1 (12.24 words) to CM2 (14.07 words), distributions largely overlapped, indicating that sentence length contributed only modestly to level differentiation.

The correlations of readability features were calculated (Figure 3). The lexical variables, especially the proportions of words in graded lexicons are strongly correlated with each other (in yellow). There are also extremely strong negative correlations (in dark blue) between the numbers of complex clauses and the proportions of words in graded lexicons. Logically, the results show a perfect positive correlation (in yellow) between the number of tokens per sentence and the number of constituents/phrases (correlation of 1). There is a strong positive correlation (in light green) between the number of words per text and the measure of lexical diversity (correlation of 0.74).

6.2. Phonemic Features

A total of 129,608 CS keys are obtained, compared with 91,786 keys reported by Gala et al. (2024) on the same Alector corpus. In that study, sentences containing more than 100 phonemes were excluded; this constraint is not applied here.

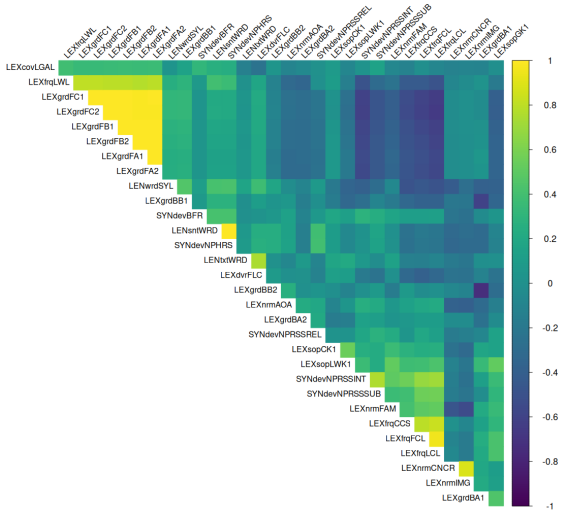


Figure 3: Correlation heatmap among the readability features on the Alector texts

Position	#	%	Vowels
Side (s)	65,625	50.60	a o œ ə ⊕
Chin (c)	22,869	17.60	ɛ u ɔ
Mouth (m)	22,354	17.20	i ɑ̃ ɔ̃
Throat (t)	15,523	12.00	e y œ̃
Cheekbone (b)	3,237	2.50	ø ɛ̃

Table 4: Total number of keys by position (vowels) in the CS annotated Alector corpus

Handshape	#	%	Conson.
5	30,596	23.61	m t f ⊕
3	30,135	23.25	s ʁ
1	20,359	15.70	p d ʒ
6	19,673	15.17	l ʃ w ɲ
2	14,289	11.02	k v z
4	9,043	6.97	b n ɥ
8	3,941	3.04	j ŋ
7	1,572	1.21	g

Table 5: Total number of keys by handshape (consonants) in the CS annotated Alector corpus

Among those 129,608 CS keys, the most frequent position for vowels is the side of the face (Table 4), as it is the placement which has the most vowels and the most frequent ones in French, it also codes the silent -e as well as the absence of vowels – e.g. the single consonant t, 5**s** in '*les autres*' (cf. Table 3).

In some French CS training programs, we start by learning all the positions of the vowels in the first lesson in order to be able to gradually learn the handshapes for the consonants: one to two per lesson by combining them with all the vowels in short words, then longer and longer ones, which use the handshapes seen in the previous lessons.

As regards the distribution of handshapes, the most frequent one is the shape 5 (Table 5), which includes many coding possibilities (along with shape 6), and it is the one learned first in training (it is the shape that allows the easiest transition to the others – the open palm –, and which codes the single vowel as '*un*', **5**t in Table 3). Conversely, shape 7 only includes the sound g (rare in French) and is the only shape with one consonant). It is always the least frequent handshape (the second to last shape seen in training). The side position and the shape 5 both encode two different key structures (V and CV clusters or C and CV clusters).

Although the handshape 8 is second to last in terms of frequency, it represents the most frequent glide j (3,940 occurrences) – followed by w (1,812), then ɥ (952). The other consonant of shape 8 (ŋ) is rare in French and typical of foreign words ('*parki**ng***', '*joggi**ng***', etc.). In Alector, there is a single instance of this glide: the English proper noun 'Grunni**ng**s' in an excerpt from Harry Potter book. Conversely, the second most frequent handshape is 3, which includes the two most commonly used consonants in French. We assume that this shape is probably one of the most complex to make for beginners, unlike 5, which is the open palm. The numbers between some positions and handshapes are extremely close. The number of positions per text is compared using ANOVA: for chin and mouth, the difference is no significant (p = 0.16), as for shapes 5 and 3 (p = 0.21). In contrast, the difference between shapes 1 and 6 is significant ($p <$ 0.05).

As the number of words per text (LENtxtWRD), the total number of CS keys per text (unnormalized PHOkey) increases markedly across text levels. Median values rose from IReST (Mdn = 249.5) to CE1 (Mdn = 454), CE2 (Mdn = 597.5), CM1 (Mdn = 728), and CM2 (Mdn = 1,002). The text with the most CS keys is a CM2 text (id_189 with 1,352 keys). Conversely, the one with the fewest keys is an IReST text (id_51 with 203 keys).

For the three possible CS key configurations, there are 66.36% of CV clusters (PHOclusterCV), and respectively 25.79% and 7.83% of C and V clusters. These ratios are close to those of a corpus of 4,143 French CS keys produced by experienced cuers on read texts (Bigi, 2023), i.e., 70.72% for CV, 20.69% for C and 8.59% for V.

The most frequent POS tags are: nouns, determiners, and verbs. The most frequent lemmas are almost all invariable (closed word classes, such as determiners, prepositions, pronouns), and are small words: one or two CS keys for '*il*' (he), or the forms of '*être*' (to be) and '*avoir*' (to have). Hapax words are indeed nouns, verbs, adjectives; they are longer words (with more CS keys), which we

assume to be more complex to (de)code due to cognitive load.

The number of transitions between the side from/to throat positions (PHOpositionST) is much more frequent than between the cheekbone from/to throat positions (PHOpositionBT), respectively with 14,651 and only 665 occurrences in the corpus. Note that the distance between side from/to throat positions is not necessarily the greatest because the possible space for these peripheral positions is large (Bigi, 2025b).

Phonemic variables are generally not strongly correlated with each other (Figure 4), except (in light green) between the number of all glides and the number of each glide (j, w, ɥ) – respectively correlations of 0.78, 0.37, 0.26 –, and between the number of CS keys and the number of syllables per word (correlation of 0.58).

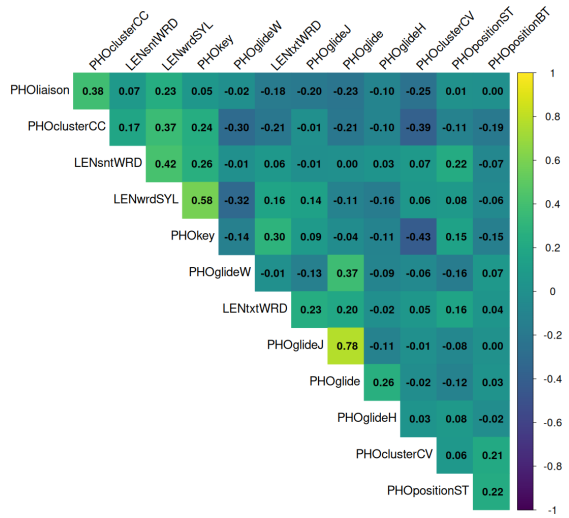


Figure 4: Correlation heatmap among the length-based variables and the phonemic variables on the Alector texts

6.3. Combining the Features

A full correlation heatmap shows that readability variables are not strongly correlated with phonemic variables as shown in Appendix B (Figure 6).

The two types of features were extracted to automatically classify texts of Alector according to their level of complexity. In total, 10 variables were automatically selected by MRMR as they provided an optimal trade-off between explanatory power and model parsimony: LENtxtWRD, LENwrdSYL, LEXdvrFLC, LEXgrdFA1, LEXnrmIMG, LEXsopCK1, SYNdevNPRSSREL, PHOkey, PHOpositionST, PHOglideJ; i.e., 7 readability variables (2 length-based, 4 lexical, 1 syntactic), and 3 phonemic variables. There are one or two variables per family in each type even if not all families are represented. In order to assess redundancy among the selected features, pairwise correlations with text level were examined (Table 6). Overall, the selected feature set reflects a balance between highly informative predictors (LENtxtWRD, LEXdvrFLC) and complementary linguistic indicators (PHOkey, PHOglideJ, LENwrdSYL, LEXgrdFA1), ensuring both predictive relevance and reduced redundancy. This supports the use of the selected variables for downstream clustering analyses.

#	Feature	Corr. with level
1	LENtxtWRD	0.907
2	LEXdvrFLC	0.712
3	PHOkey	0.429
4	PHOglideJ	0.246
5	LENwrdSYL	0.238
6	LEXgrdFA1	0.208
7	LEXsopCK1	0.160
8	PHOpositionST	0.145
9	SYNdevNPRSSREL	-0.0118
10	LEXnrmIMG	-0.286

Table 6: Selected features and their correlation with the text level

The relevant number of clusters suggested by the K-means algorithm applied on our data is 3. The distribution of text levels across clusters reveals a clear gradient of text difficulty (Figure 5). Cluster 1 is dominated by beginner level texts (IReST, CE1), whereas cluster 3 is enriched in advanced level texts (CM1, CM2). Cluster 2 shows a more heterogeneous composition, encompassing primarily intermediate levels texts (CE1, CE2, CM1). Although cluster boundaries do not perfectly align with the predefined text levels, the observed distribution suggests that the clustering captures latent dimensions of complexity. We proposed an initial model for selecting texts to be recorded in CS based on different features. However, the target variable is the text level. These levels are intended for children learning French, but not French CS. This calls into question whether the model actually captures French CS learning difficulty. We acknowledge this limitation. Further research is needed to verify the impact of the text level.

7. Conclusion and Discussion

This paper presented the extraction of textual and phonemic features to automatically classify texts for French CS and reading training according to their complexity level. The creation of freely accessible resources in CS addresses an important societal need aimed to improving the inclusion and comprehension of people with hearing loss.

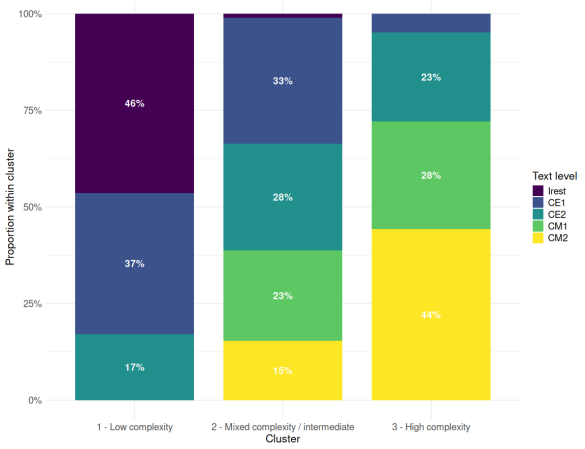


Figure 5: Proportional distribution of text levels within each cluster

In total, 41 variables potentially linked to CS complexity were inspected. Several variable types were considered, including both textual- and phoneme-level measures, which are rarely examined jointly. Because the literature on complexity variables in French CS remains limited, some hypotheses cannot yet be verified within the scope of this paper. The results showed positive correlations between readability variables, especially the length-based and lexical. The phonemic features are poorly correlated with each other and with readability variables, except the number of CS keys, glides or syllables per words. The analysis of key frequency appears consistent with previous studies (Bigi, 2023; Gala et al., 2024), with reported French syllable frequencies, and with the key learning order commonly taught; further surveys of the target audience are needed to better document current CS teaching practices.

Modeling complexity in terms of readability or listenability traditionally relies on the use of data evaluated by the target audience, which we didn't have. Therefore, we proposed an initial exploratory unsupervised model to classify texts that will be generated in French CS. We identified, extracted and combined readability and phonemic features. Future work includes choosing texts and words from Alector corpus to be recorded. The videos will then be annotated with a CS coding hand using the SPPAS tool, before being evaluated by deaf or hard-of-hearing CS users. The learning complexity of the recorded texts will be assessed to refine the features. This evaluation will allow to propose customization options for users of the learning platform, such as adding specific filters (by CS keys, text length, with or without glides, etc.).

Additional factors may also contribute to CS complexity and deserve further investigation, including fluency-related measures – e.g., syllables per minute, pause duration, pauses per minute as in Ozawa et al. (2024) –, and learner-specific features. Further work could also incorporate phoneme-grapheme consistency criteria that may complicate (de)coding in French CS, such as graphemes with context-dependent pronunciations governed by more or less regular rules.

Lay Summary

In this work, we analyze a corpus of texts that will be translated into Cued Speech. Cued Speech is used with deaf and hard-of-hearing people to improve their understanding of spoken language. This communication mode combines gestures and speech. There are several hand movements in different positions around the face to represent the sounds. Although this method is known to be helpful, there are still not many tools available to learn Cued Speech, especially in French. In order to create learning videos, we propose to extract and analyze 41 textual and phonemic features that might be more complex in French Cued Speech. We mainly used two tools: FABRA and SPPAS. The results show some correlations. A first model is proposed for selecting texts to be recorded for learning French Cued Speech.

Acknowledgments

This work received support from the French government under the France 2030 investment plan, as part of the *Initiative d'Excellence d'Aix Marseille Université* - AMIDEX (AMX-22-RE-AB-022), VizLector project. It is also a part of the Automatic Cued Speech project (APa2022_022), supported by FIRAH (*Fondation Internationale de la Recherche Appliquée sur le Handicap*, International Foundation of Applied Disability Research). Finally, the authors thank the reviewers of the first version of the paper and the members of the French ALPC association for the LfPC lessons.

8. Bibliographical References

Mélody Artaz. 2011. L'apprentissage de la Langue française Parlée Completée par les enfants sourds d'âge scolaire : Conception d'un loto LPC. Master's thesis, Université Stendhal.

Brigitte Bigi. 2015. SPPAS - Multi-Lingual Approaches to the Automatic Annotation of Speech. *The Phonetician. Journal of the International Society of Phonetic Sciences*, 111:54–69.

Brigitte Bigi. 2023. An analysis of produced versus predicted French Cued Speech keys. In *10th*

Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, pages 24–28, Poznań, Poland.

Brigitte Bigi. 2025a. Bridging the Gap: Design and Evaluation of an Automated System for French Cued Speech. In *Proceedings of the 8th International Conference on Natural Language and Speech Processing (ICNLSP-2025)*, pages 8–18, Southern Denmark University, Odense, Denmark. Association for Computational Linguistics.

Brigitte Bigi. 2025b. Spatial Analysis of Hand Positions in French Cued Speech (LfPC). In *16th International Conference on Linguistic Research and Applications*, Paris, France.

Malika Charrad, Nadia Ghazzali, Véronique Boiteau, and Azam Niknafs. 2014. NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software*, 61(6):1–36.

R. Orin Cornett. 1967. Cued Speech. *American Annals of the Deaf*, 112(1):3–13.

Nicolas De Jay, Simon Papillon-Cavanagh, Catharina Olsen, Nehme El-Hachem, Gianluca Bontempi, and Benjamin Haibe-Kains. 2013. mRMRe: an R package for parallelized mRMR ensemble feature selection. *Bioinformatics*, 29(18):2365–2368.

Chris Ding and Hanchuan Peng. 2005. Minimum Redundancy Feature Selection from Microarray Gene Expression Data. *Journal of Bioinformatics and Computational Biology*, 3(2):185–205.

Rudolph Flesch. 1948. A New Readability Yardstick. *Journal of Applied Psychology*, 32(3):221–233.

Núria Gala, Brigitte Bigi, and Marie Bauer. 2024. Automatically Estimating Textual and Phonemic Complexity for Cued Speech: How to See the Sounds from French Texts. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1817–1824, Torino, Italia. ELRA and ICCL.

Ludivine Javourey-Drevet, Stéphane Dufau, Thomas François, Núria Gala, Jacques Ginestié, and Johannes C. Ziegler. 2022. Simplification of literary and scientific texts to improve reading fluency and comprehension in beginning readers of French. *Applied Psycholinguistics*, 43(2):485–512.

Elaine Jones, Robert Strom, and Susan Daniels. 1989. Evaluating the Success of Deaf Parents. *American Annals of the Deaf*, 134(5):312–316.

Liliane Kandel and Abraham Moles. 1958. Application de l'indice de Flesch à la langue française. *Cahiers Etudes de Radio-Télévision*, 19:253–274.

András Kornai. 2007. *Mathematical Linguistics*. Springer Science & Business Media.

Katsunori Kotani, Shota Ueda, Takehiko Yoshimi, and Hiroaki Nanjo. 2014. A Listenability Measuring Method for an Adaptive Computer-assisted Language Learning and Teaching System. In *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing*, pages 387–394.

Katsunori Kotani and Takehiko Yoshimi. 2017. Effectiveness of Linguistic and Learner Features for Listenability Measurement Using a Decision Tree Classifier. *The Journal of Information and Systems in Education*, 16(1):7–11.

Mélanie Lancien and Brigitte Bigi. 2025. French Cued Speech rhythm: first findings on the relationship between hand position and segments' duration. In *11th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science, Linguistics and Low Resourced Languages*, Poznań, Poland.

Manuel Leitao, Elodie Venti, Thomas Sigiez, Christophe Laroche, Marie Perini, and Agnès Piquard-Kipffer. 2021. Projet LogilecSur: quelles stratégies enseignantes pour guider des élèves sourds vers l'autonomie en compréhension écrite? In *IDEKI 2021 - 4ème colloque international Didactiques et métiers de l'humain*, Pont-à-Mousson, France.

Jacqueline Leybaert and Carol J. LaSasso. 2010. Cued Speech for Enhancing Speech Perception and First Language Development of Children With Cochlear Implants. *Trends in Amplification*, 14(2):96–112.

Aïnizé Olhagaray. 2013. « Le Pirate Codeur » : Élaboration d'un matériel ludique visant à entraîner et automatiser le décodage de la Langue française Parlée Complétée (LPC) : à destination des enfants sourds ayant reçu une introduction tardive du code LPC. Master's thesis, Université Bordeaux Segalen.

Lucía Ormaechea and Nikos Tsourakis. 2024. Automatic text simplification for French: model fine-tuning for simplicity assessment and simpler text generation. *International Journal of Speech Technology*, 27(4):957–976.

Minami Ozawa, Rodrigo Wilkens, Kaori Sugiyama, and Thomas François. 2024. Modéliser la facilité d'écoute en FLE: vaut-il mieux lire la transcription ou écouter le signal vocal ? In

35èmes Journées d'études sur la Parole (JEP 2024) 31ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2024) 26ème Rencontre des étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL 2024), volume 1, pages 549–566. ATALA & AFPC.

Agnès Piquard-Kipffer. 2016. Un album numérique pour raconter une histoire avec un avatar narrateur. In *XVIèmes rencontres internationales en orthophonie - Orthophonie et technologies innovantes*, Paris, France.

R Core Team. 2021. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*.

Stéphane Rauzy, Grégoire Montcheuil, and Philippe Blache. 2014. MarsaTag, a tagger for French written texts and speech transcriptions. In *Second Asian Pacific Corpus linguistics Conference*, pages 220–220, Hong Kong, China.

Valérie Sabbagh. 2012a. *Le Petit Clown 2 Le LPC pour les enfants : Entraînement et perfectionnement*. ALPC, Paris, France.

Valérie Sabbagh. 2012b. *Le Petit Clown Le LPC pour les enfants : L'imagier d'Agathe*. ALPC, Paris, France.

Sanjana Sankar. 2024. *Automatic recognition and generation of French Cued Speech using deep learning*. Ph.D. thesis, Université Grenoble Alpes.

Rodrigo Wilkens, David Alfter, Xiaoou Wang, Alice Pintard, Anaïs Tack, Kevin P Yancey, and Thomas François. 2022. FABRA: French Aggregator-Based Readability Assessment toolkit. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1217–1233, Marseille, France. ELRA.

World Health Organization. 2021. *World Report on Hearing*. World Health Organization. ISBN: 978-92-4-002048-1.

9. Language Resource References

Brigitte Bigi, Maryvonne Zimmermann, and Carine André. 2022. CLeLfPC: a Large Open Multi-Speaker Corpus of French Cued Speech. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 987–994, Marseille, France. ELRA.

Gala, Núria and Tack, Anaïs and Javourey-Drevet, Ludivine and François, Thomas and Ziegler, Johannes C. 2020. *Alector: A Parallel Corpus of Simplified French Texts with Alignments of Misreadings by Poor and Dyslexic Readers*. ELRA.

Vital-Durand, François. 2011. *International Reading Speed Texts IResT (French version)*.

A. Overview of all the Variables

All details and references on readability variables from FABRA can be found in Wilkens et al. (2022).

B. Correlation Matrix

Correlations among all variables are shown as percentages for readability (Figure 6).

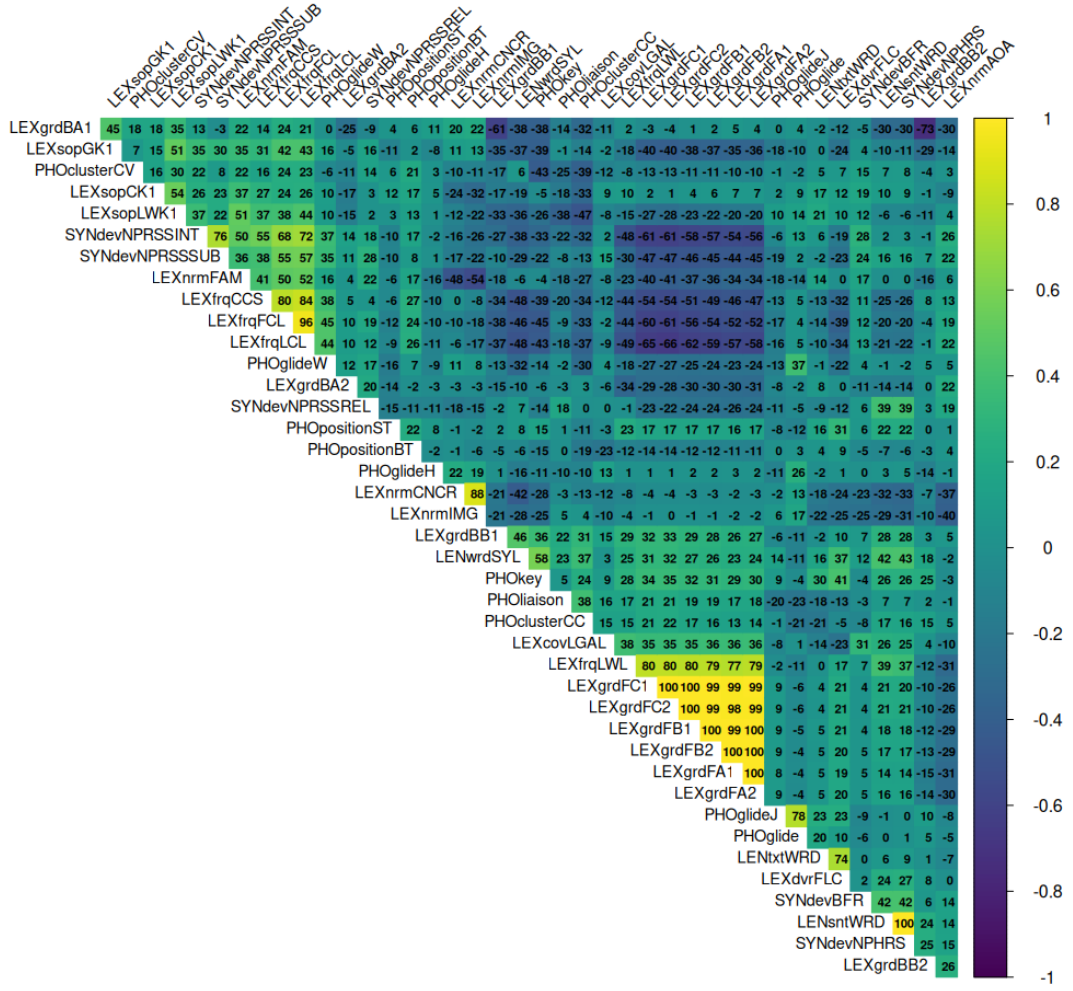


Figure 6: Correlation heatmap among the readability and phonemic features on the Alector texts

	Variable	Description
Length based		
Word length	LENwrdSYL	Number of syllables per word
Sentence length	LENsntWRD	Number of tokens per sentence, excluding punctuation
Text length	LENtxtWRD	Number of words per text, excluding punctuation; unnormalized
Lexical Variables		
Content overlap	LEXcovLGAL	Any lemma is shared in any sentences
Lexical diversity	LEXdvrFLC	CTTR of all types of lemma forms of nouns, proper nouns, verbs, adjectives and adverbs in the text, considering all tokens
Lexical frequency	LEXfrqCCS	Frequency of surface form of all nouns, proper nouns, verbs, adjectives and adverbs based on the occurrence at CHILDES corpus
Lexical frequency	LEXfrqFCL	Frequency of lemma form of all nouns, proper nouns, verbs, adjectives and adverbs based on the occurrence at FLELex corpus
Lexical frequency	LEXfrqLCL	Frequency of lemma form of all nouns, proper nouns, verbs, adjectives and adverbs based on the occurrence at Lexique3 corpus
Lexical frequency	LEXfrqLWL	Frequency of lemma form of all words based on the occurrence at Lexique3 corpus
Graded lexicons	LEXgrd[BA1/BA2/BB1/BB2]	Proportion of words in Beacco's French Reference Level Descriptors for each CEFR level (A1 to B2)
Graded lexicons	LEXgrd[FA1/FA2/FB1/FB2 FC1/FC2]	Frequency of words in FLELex resource for each CEFR level (A1 to C2)
Lexical norms	LEXnrmAOA	Age of acquisition of each word
Lexical norms	LEXnrmCNCR	Words level of concreteness
Lexical norms	LEXnrmFAM	Words familiarity, also called subjective frequency
Lexical norms	LEXnrmIMG	Imageability of each word
Lexical sophistication	LEXsopCK1	Number of words in the first frequency bands of 1,000 words of CHILDES
Lexical sophistication	LEXsopGK1	Number of words in the first frequency bands of 1,000 words of Gougenheim vocabulary list
Lexical sophistication	LEXsopLWK1	Number of surface form words in the first frequency bands of 1,000 words of Lexique3
Syntactic Variables		
Language development	SYNdevBFR	Number of words before the main verb
Language development	SYNdevNPHRS	Number of constituents/phrases
Language development	SYNdevNPRSSINT	Number of different types of phrases/ constituents of type SINT (internal conjugate clause - *proposition conjuguée interne*)
Language development	SYNdevNPRSSREL	Number of different types of phrases/ constituents of type SREL (relative clause - *proposition relative*)
Language development	SYNdevNPRSSSUB	Number of different types of phrases/ constituents of type SSUB (subordinate clause - *proposition subordonnée*)

Table 7: Readability variables description

	Variable	Description
CS key frequency	PHOkey	Number of CS keys, (un)normalized per character
CS key frequency	PHOpositionST	Number of transitions between side and throat positions (and conversely); normalized per PHOkey-1
CS key frequency	PHOpositionBT	Number of transitions between cheekbone and throat positions (and conversely); normalized per PHOkey-1
CS key frequency	PHOclusterCV	Number of consonant/vowel clusters; normalized per PHOkey
Phonemic frequency	PHOglideW	Number of glides (semi-consonants or semi-vowels) w; normalized per PHOkey
Phonemic frequency	PHOglideH	Number of glides (semi-consonants or semi-vowels) ɥ; normalized per PHOkey
Phonemic frequency	PHOglideJ	Number of glides (semi-consonants or semi-vowels) j; normalized per PHOkey
Phonemic frequency	PHOglide	Number of glides (semi-consonants or semi-vowels) w, ɥ, j; normalized per PHOkey
Phonemic frequency	PHOliaison	Number of obligatory liaisons; normalized per LENtxtWRD-1
Phonemic frequency	PHOclusterCC	Number of double consonants; normalized per LENtxtWRD

Table 8: Phonemic variables description

Can LLMs Control Readability? A Multi-Dimensional Evaluation Framework for CEFR-Controlled Arabic Generation

Nour Rabih, Chatrine Qwaider, Ted Briscoe
Mohamed bin Zayed University of Artificial Intelligence (MBZUAI)
{nour.rabih, chatrine.qwaider, ted.briscoe}@mbzuai.ac.ae

Abstract

While Large Language Models (LLMs) can generate fluent Arabic text, their ability to reliably control readability levels remains unclear. We propose a multi-dimensional evaluation framework for Common European Framework of Reference for Language (CEFR)-controlled Arabic text generation, assessing whether instruction-following LLMs can serve as reliable generators for adaptive language learning. Our framework integrates controlled prompting, automatic readability prediction using a validated Taha-19 model, lexical constraint validation, and syntactic complexity profiling. Results show that structured prompting substantially improves CEFR alignment. In particular, CEFR-guided prompting with lexical constraints achieves the highest conformity to reference linguistic profiles (0.91 cosine similarity) and near-perfect agreement with predicted readability levels (0.99), while unconstrained prompting exhibits weak control. These findings establish an empirical foundation for integrating readability-aware Arabic text generation into adaptive educational systems.

Keywords: Arabic Text Generation, Readability Assessment, Large Language Models, CEFR-controlled Readability, Adaptive Language Learning

1. Introduction

Large Language Models (LLMs) are increasingly used in educational technologies to generate reading materials, exercises, and language-learning content on demand. In such systems, controlling the linguistic difficulty of generated text is essential, as the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001), which defines proficiency levels from A1 to C2, is widely adopted in educational settings. For safe deployment in learning environments, generative models must therefore produce not only fluent text, but content that is reliably aligned with learners' reading abilities. This challenge is particularly acute for Arabic. Its rich morphology, flexible word order, and syntactic variability make readability control substantially harder than in many high-resource languages. However, despite growing interest in readability-controlled generation (Ribeiro et al., 2023; Trott and Rivière, 2024), most existing work is based on prompting a LLM and relies on a single readability metric. This limits its applicability to adaptive learning systems, where generated content should be continuously assessed and adjusted in a fine-grained way to individual learners.

In this work, we propose a multi-dimensional framework for CEFR-controlled Arabic text generation with LLMs, designed for integration into adaptive educational pipelines. Our framework combines controlled prompting, automatic readability prediction, CEFR-aligned lexical constraints, and syntactic complexity analysis to evaluate whether LLMs can generate Arabic texts that are fluent and pedagogically appropriate across proficiency levels. We first establish a data-driven alignment between CEFR levels and the fine-grained 19-level Taha-Thomure (2017) readability scale by applying the pretrained BAREC readability model (Elmadani et al., 2025) to CEFR-labeled learner essays from the ZEABUC and ARWI Arabic Essay Scoring datasest (Habash and Palfreyman, 2022; Qwaider et al., 2025). In addition, we construct detailed linguistic profiles for each CEFR level based on these corpora, capturing characteristic syntactic and lexical properties. These empirically derived level-specific profiles are later used as evaluation benchmarks to systematically assess the structural and lexical alignment of generated essays with target proficiency levels.

We then prompt GPT-4o[1] to generate Arabic essays under different conditions, including prompts specifying only the target CEFR level and prompts augmented with CEFR-appropriate lexical and syntactic constraints. The outputs are evaluated using the BAREC model for fine-grained readability estimation and assessed against the constructed reference profiles to determine the extent to which their lexical and syntactic properties align with the expected specifications of each CEFR level.

Our contributions are threefold:

- The first multi-dimensional study of CEFR-controlled Arabic text generation with LLMs,

[1] https://openai.com/index/hello-gpt-4o/

Proceedings of the Joint Workshop on Readability and Text Simplification (READIxTSAR) @ LREC 2026, pages 74–88
11 May 2026.

framed within an adaptive learning perspective.

- A principled alignment between CEFR and a fine-grained Arabic readability scale, enabling continuous modeling of text difficulty.
- A dataset of CEFR-controlled Arabic essays[2] to support future research on personalized, readability-aware text generation for Arabic.

The rest of the paper is structured as follows. §2 reviews reviews related work. §3 presents the datasets used.§4 describes the proposed evaluation framework, §5 details the experimental setup. §6 reports the results and analysis across prompting conditions. §7, concludes the paper and outlines directions for future work.

2. Related Work

Arabic Readability Research on Arabic readability has explored a range of standards, datasets, and modeling approaches. Early work by Khallaf and Sharoff (2021) modeled Arabic readability using a CEFR-inspired scheme with a coarse three-level classification. Other studies have adopted grade-based readability levels for first-language (L1) readers or instructional proficiency bands for second-language (L2) learners, reflecting curriculum-driven rather than standardized assessment criteria (Cavalli-Sforza et al., 2018). More recent dataset-driven efforts, such as DARES, provide both coarse-grained(4 levels) and fine-grained (12 levels) readability annotations, enabling more detailed modeling of Arabic text complexity (El-Haj et al., 2024). In parallel, the BAREC dataset (Elmadani et al., 2025) offers a fine-grained 19-level readability scale (Taha-19) designed specifically for Arabic, a pedagogically motivated framework inspired by Taha-Thomure (2017), and is accompanied by a fine-tuned transformer-based model for automatic readability prediction. Other transformer-based approaches have further advanced readability modeling, Readme++ introduces a multilingual CEFR classification model applicable to Arabic (Naous et al., 2024), however, its performance on Arabic is lower compared to higher-resource languages, highlighting the additional challenges associated with Arabic readability modeling. In this work, we leverage BAREC for fine-grained readability assessment, enabling evaluation of CEFR alignment and linguistic complexity in generated Arabic text.

Controlled Text Generation and Readability Evaluation Prior work has investigated readability-controlled summarization and simplification, often using traditional readability formulas such as Flesch–Kincaid or Lexile scores (Ribeiro et al., 2023; Trott and Rivière, 2024). More recent studies have examined whether instruction-tuned LLMs can follow explicit readability constraints specified in prompts, evaluating generated text using trained SVMs from Xia et al. (2016) to predict the CEFR level of the generated essay. Most existing analyses focus on English and rely on a single readability metric, limiting their ability to capture nuanced linguistic variation. Work on Arabic is especially scarce, and to our knowledge, no prior study has systematically evaluated CEFR-controlled Arabic text generation, nor examined syntactic complexity as an additional diagnostic signal.

Level	Arabic Prompt	English Translation
Beginner	اوصف يومك المفضل	Describe your favorite day.
Interm.	كيف يمكننا التعامل مع الضغوط النفسية؟	How can we deal with psychological stress?
Advanced	تحدث عن أهمية التعليم الرقمي في عصرنا الحالي.	Talk about the importance of digital education in our current era.

Table 1: Example CEFR-aligned essay prompts

3. Data

We build our generation experiments and evaluation on two complementary data sources:

ZAEBUC.[3] The Zayed Arabic-English Bilingual Undergraduate Corpus (ZAEBUC) (Habash and Palfreyman, 2022), is a publicly available dataset of short Arabic essays written by first-year university students and annotated for linguistic features including CEFR ratings. it provides 214 manually corrected Arabic essays with consistent CEFR labels, offering a reliable basis for studying proficiency.

ARWI.[4] The Arabic Read Write and Improve dataset (Qwaider et al., 2025), is a publicly available dataset that contains synthetic generated essays for automated essay scoring . From this work, we use both the synthetic essays (3220 essays) and their associated leveled prompts and topics, which specify target proficiency levels. A few examples are shown in Table 1. In our experiments, these prompts and topics are reused to generate new essays using GPT-4o. The original dataset organizes prompts into three broad proficiency categories: Beginner, Intermediate, and Advanced. To

[2] https://github.com/noorrabih/CEFR-Controlled-Arabic-Generation-Data.git

[3] https://sites.google.com/view/zaebuc/home

[4] https://github.com/mbzuai-nlp/arabic-aes-bea25

align with the CEFR framework, we map these categories into six CEFR levels as follows: Beginner → A1–A2, Intermediate → B1–B2, and Advanced → C1–C2. The dataset contains 161 prompts in total, distributed as 50 Beginner, 56 Intermediate, and 55 Advanced prompts.

We merge ARWI and ZAEBUC to create a unified proficiency-based essay collection. These datasets are selected because they are annotated and graded according to students' writing proficiency levels, making them suitable for modeling learner ability. By combining them, we construct coherent proficiency profiles aligned with CEFR levels. Table 2 presents the distribution of essays after merging the datasets. The *Unassessable* essays from ZAEBUC (6 instances) are excluded to ensure consistency across proficiency levels.

CEFR Level	ARWI	ZAEBUC	Total
A1	500	0	500
A2	500	7	507
B1	560	110	670
B2	560	80	640
C1	550	11	561
C2	550	0	550
Unassessable	0	6	6

Table 2: Distribution of essays across CEFR levels for ARWI and ZAEBUC datasets.

SAMER Lexicon We also utilize the SAMER readability lexicon (Al Khalil et al., 2020), a large-scale lexical resource for Modern Standard Arabic containing 26,578 manually annotated lemmas extracted from news and literary corpora. Each lemma is assigned one of five grade-based readability levels, annotated in triplicate by language professionals from different Arab regions. The lexicon combines corpus frequency information with expert judgment, providing a reliable lexical complexity signal that we incorporate into our proficiency modeling. To align SAMER with CEFR, we map its five levels to CEFR bands as follows: A1–A2 → Level 1, B1 → Level 2, B2 → Level 3, C1 → Level 4, and C2 → Level 5. This mapping preserves the ordinal progression of lexical complexity across both frameworks and enables consistent vocabulary constraints during controlled generation.[5]

[5]The mapping from SAMER levels to CEFR bands is heuristic and based on aligning the ordinal progression of lexical difficulty across the two frameworks. While this provides a practical approximation for controlled generation, more principled alignment strategies (e.g., data-driven calibration or joint modeling) remain an interesting direction for future work.

4. Methodology

4.1. Assessment Standards

We adopt two complementary readability standards to evaluate the generated texts: the **CEFR** and the fine-grained **Taha-19** readability scale. Together, these frameworks enable both coarse proficiency assessment and detailed analysis of linguistic complexity.

CEFR. The CEFR (Council of Europe, 2001) defines six proficiency levels (A1–C2) and is widely used to describe non-native learners' linguistic attainment across reading, writing, listening, and speaking. It provides a standardized framework for assessing and comparing language proficiency across different educational contexts.

Taha-19. The Taha-19 scale is a fine-grained Arabic readability framework consisting of 19 levels, pedagogically motivated and derived from the Taha–Thomure model (Taha-Thomure, 2017). It defines 19 ordered levels of reading difficulty tailored to Arabic, capturing fine-grained progression in linguistic complexity beyond coarse CEFR categories. Lower levels correspond to simple sentence structures, limited vocabulary, and minimal morphological variation, while higher levels reflect increased syntactic embedding, richer vocabulary, and more complex discourse structures. It is accompanied by the **BAREC** transformer-based predictor, which is trained on a large, balanced Arabic readability corpus. We use the BAREC-model to assign sentence-level Taha-19 readability scores.

Taha-19 and CEFR Alignment. To enable joint analysis across coarse CEFR levels and fine-grained readability, we establish an empirical alignment between the two scales using the ZAEBUC and ARWI corpora. Taha-19 readability scores are inferred using the BAREC model.[6] Each essay is segmented into sentences using standard Arabic punctuation-based splitting (full stops, question marks, and exclamation marks). Sentence-level readability is inferred for each segment, and document-level readability is computed as the average of sentence-level predictions, yielding a continuous fine-grained score for every CEFR-labeled essay.

Figure 1 shows the distribution of Taha-19 readability levels across CEFR Labels. A clear trend is observed: A1–A2 texts exhibit the highest density in lower readability bands (ranging betweend 6–10), B1–B2 texts concentrate in intermediate ranges (ranging between 10–13), and C1–C2 texts

[6]https://huggingface.co/CAMeL-Lab/readability-arabertv2-d3tok-CE

Figure 1: Taha-19 readability scores alignment with CEFR levels.

align with higher bands (ranging between 12–14). This progression is further confirmed by the mean Taha-19 scores computed for each CEFR level: A1 (7.61) and A2 (9.26) correspond to lower readability bands, B1 (11.03) and B2 (11.81) fall within intermediate ranges, while C1 (12.78) and C2 (12.93) reflect higher readability levels. The steady increase in mean scores across CEFR levels demonstrates strong alignment between the Taha-19 scale and CEFR proficiency progression. To statistically quantify this alignment, we compute the Spearman rank correlation between ordinal CEFR levels (A1–C2 mapped to 1–6) and document-level Taha-19 scores. The results show a strong positive correlation ($\rho = 0.84$, $p < 0.001$), indicating a significant monotonic relationship between CEFR proficiency progression and fine-grained readability.

4.2. Essay Generation

We formulate the generation task as producing Arabic essays on specified topics, enabling controlled evaluation of readability across CEFR levels. This setup reflects real-world usage, where educators and content creators rely on LLMs, such as GPT models, to generate pedagogical content aligned with learner proficiency levels (Kasneci et al., 2023; Whalen et al., 2023).

To examine whether LLMs can respond to increasingly explicit readability instructions, we design five instructional prompting styles that incrementally encode more information about the target reading level.

We consider five prompting conditions with increasing degrees of control. P1 provides no explicit readability information and serves as an unconstrained baseline. P2 specifies only the target CEFR level. P3 introduces syntactic control by imposing CEFR-aligned constraints on sentence structure and complexity based on the created profiles (Section 4.3). P4 augments level specification with CEFR-aligned vocabulary constraints by requiring the inclusion of level-appropriate words.(Section 4.4). Finally, P5 combines both vocabulary and syntactic constraints. The prompts used for each condition along with their translations are provided in Appendix D.

Due to the structure of the ARWI dataset prompts (Section 3), P1 is evaluated at three broad proficiency bands: A, B, and C. In contrast, P4 and P5 are evaluated at five finer-grained CEFR levels (A, B1, B2, C1, and C2), reflecting the granularity supported by the SAMER vocabulary classification system, which categorizes words into five proficiency levels.

4.3. Linguistic profile Construction

Syntax To construct syntactic profiles for each CEFR level, we conduct an in-depth syntactic analysis on the data. Each essay is processed using the CamelParser[7] (Elshabrawy et al., 2023), from which we extract three categories of features: (1) Part-of-Speech (POS) tags, (2) dependency relations, and (3) combined dependency-POS patterns.

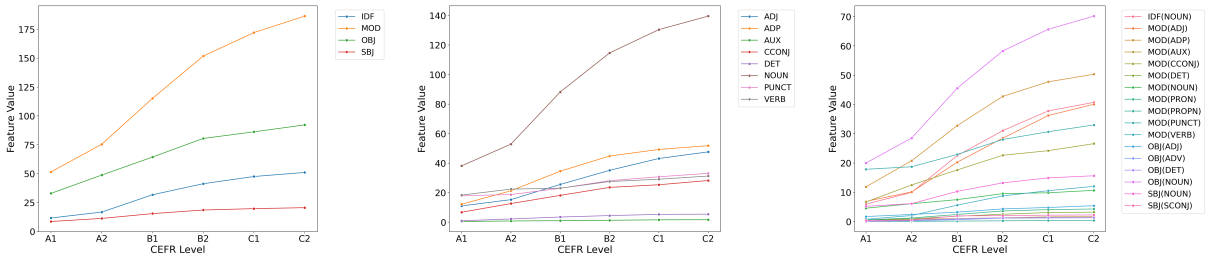
The analysis yields a total of 15 unique POS tags, 7 dependency relations, and 72 dependency–POS combinations. We then examine the distribution of these features across CEFR levels to identify meaningful and consistent trends, as illustrated in Figure 2(a,b,c).

Across these subfigures, features demonstrate clear monotonic or near-monotonic increases with proficiency level, indicating structural and syntactic development. Based on the consistency of these trends, we select 29 syntactic patterns that exhibit clear developmental progression and incorporate them into the level-specific syntactic profiles.

Dependency Tree Depth In addition to categorical syntactic features, we analyze structural complexity through dependency tree depth. For each sentence, we compute the depth of its dependency tree and then calculate the mean depth per essay. This aggregated measure serves as an indicator of syntactic embedding and hierarchical complexity across proficiency levels.

The distribution of average essay dependency depths across CEFR levels is illustrated in Figure 3. As shown in the figure, higher proficiency levels tend to exhibit greater tree depths, reflecting increased structural embedding and more complex hierarchical constructions. This trend is quantitatively confirmed by the mean dependency depth values for each CEFR level: A1 (4.90) and A2 (7.12) show comparatively shallow syntactic structures, B1 (8.97) and B2 (9.71) demonstrate moder-

[7] https://github.com/CAMeL-Lab/camel_parser, version 2.0



(a) Dependency (b) POS (c) dep(POS)

Figure 2: Mean of Syntactic features across CEFR levels. (a) shows the progression of dependency relations, (b) presents the distribution of POS tags, and (c) depicts the combined dependency–POS patterns.

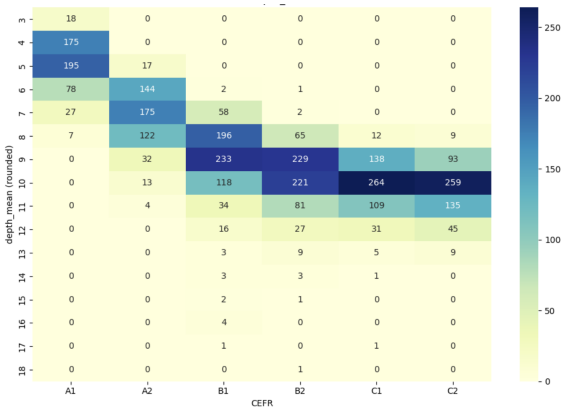


Figure 3: Distribution of mean dependency tree depths across CEFR levels

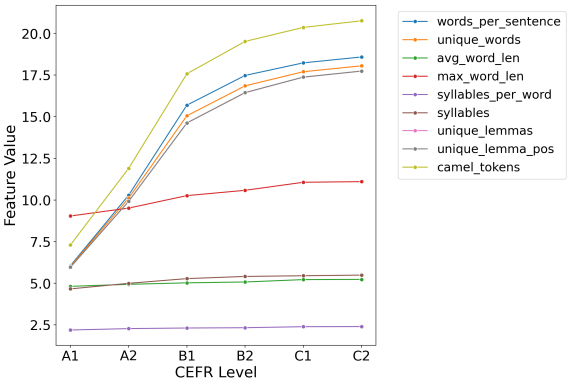


Figure 4: Lexical and surface-level features across CEFR levels.

ate structural expansion, while C1 (10.05) and C2 (10.25) exhibit the deepest hierarchical constructions. The steady increase in mean depth across proficiency levels suggests that syntactic embedding grows progressively with CEFR advancement, supporting the role of dependency depth as an indicator of linguistic complexity.

Lexical and Surface-Level In addition to syntactic profiling, we extract lexical and surface-level features to capture vocabulary richness, morphological variation, and overall textual complexity across CEFR levels. These features complement the syntactic rule set by providing quantitative indicators of lexical development.

We compute core lexical statistics, including the total number of words, number of unique words, and average sentence length (words per sentence). To measure lexical diversity beyond surface forms, we calculate the number of unique lemmas and unique lemma–POS pairs, enabling analysis of conceptual vocabulary breadth and syntactic-functional variation.

We further model word-level complexity using length and syllable-based features, including average and maximum word length, as well as average and maximum syllable counts per word. These measures capture phonological and morphological complexity associated with proficiency progression. As shown in Figure 4, these lexical and surface-level features exhibit a clear increasing trend across CEFR levels, reflecting systematic growth in lexical diversity, sentence length, and morphological complexity as proficiency advances from A1 to C2.

Together, these syntactic, structural, and lexical features define comprehensive proficiency-specific linguistic profiles, capturing both categorical and gradient aspects of language development. These profiles serve as reference benchmarks for evaluating generated texts and modeling CEFR-aligned complexity. A detailed summary of feature statistics across CEFR levels is provided in Appendix A.

4.4. Vocabulary List Construction

To support lexical control in the constrained prompting condition, we automatically construct CEFR-aligned vocabulary lists for each topic. For a given topic and target CEFR level, we first prompt GPT to generate a candidate set of words using the template below:

> You are an Arabic vocabulary instructor designing word lists aligned with CEFR levels.
> Target CEFR level: {cefr}

Topic (Arabic): "{topic}"
Task:

- Suggest a list of vocabulary suitable for a learner at level {cefr}, related to the topic above, that the learner can use to write an Arabic essay.
- Ensure the words are appropriate in frequency and difficulty for {cefr}.
- Output only the list of Arabic words, separated by commas.
- Number of words: between {min} and {max}.

Example format (no numbering, no extra text):
`word1, word2, word3, ...`

The generated candidate lists are then filtered using **SAMER** (Al Khalil et al., 2020), a five-level Arabic lexical difficulty lexicon. Only words whose SAMER level matches the target CEFR band are retained.

If the filtered list contains fewer than the minimum required number of words (3 words), we repeat the generation-filtering process. If the list remains insufficient after multiple attempts, we apply a vocabulary relevance model (Reimers and Gurevych, 2019) that ranks SAMER words by semantic similarity to the topic and selects the highest-scoring items from the appropriate difficulty band. This fallback mechanism ensures that each topic is associated with a sufficient number of level-appropriate and semantically relevant lexical constraints.

A complete example of the constructed prompts and their corresponding CEFR-aligned vocabulary lists is provided in Appendix C.

4.5. Evaluation Metrics

To evaluate how closely the generated readability-controlled texts match the reference CEFR-level profiles, we compute the cosine similarity between their feature vectors. Specifically, given two real-valued feature vectors P_i (the CEFR reference profile) and Q_i (the generated text), the cosine similarity is calculated as:

$$\cos(\theta) = \frac{\sum_i P_i Q_i}{\sqrt{\sum_i P_i^2} \cdot \sqrt{\sum_i Q_i^2}} \quad (1)$$

Cosine similarity is particularly suitable in our setting because it measures the agreement between high-dimensional feature representations (over 40 linguistic features) in terms of their relative distributional patterns rather than absolute magnitudes. This is important for readability profiling, where we aim to capture how features co-vary across CEFR levels, rather than their raw values. Compared to alternative measures such as Euclidean distance, which is sensitive to scale and absolute differences in individual features, cosine similarity focuses on the orientation of feature vectors in the shared space. This makes it more robust when combining heterogeneous linguistic features (e.g., counts, ratios, and averages) and better reflects whether the generated profiles follow the same structural trends as the reference CEFR profiles.

Before computing cosine similarity, we normalize all features using standard score normalization (z-score scaling) to ensure that features with larger numerical ranges do not dominate the similarity computation.

Evaluation Granularity To provide a comprehensive analysis, we compute cosine similarity at multiple levels of granularity:

- **Overall**: across all levels and features jointly.
- **Per-cluster**: we group the profile features into 4 groups and evaluate each separately. Table 3 summarizes the feature groupings used in our analysis.
- **Per-level**: across all features within each CEFR level.
- **Per-feature**: across CEFR levels for each individual feature.

5. Experimental Setup

5.1. Vocabulary List Construction

Generation. For each prompt–CEFR pair, we generate candidate vocabulary lists using GPT-4o with temperature set to 0.6 to balance diversity and control. The number of requested words is CEFR-dependent: A1 (20–30), A2 (25–35), B1 (30–40), B2 (35–45), C1 (40–50), and C2 (40–60). The model is instructed to output comma-separated Arabic words only, without explanations or formatting.

SAMER Validation and Morphological Alignment. All generated words are morphologically analysed using the CAMeL Tools [8] (Obeid et al., 2020) to obtain lemma–POS representations. These are matched against the SAMER readability lexicon to verify that each word belongs to the intended difficulty band. Words whose SAMER readability does not match the target level are discarded. For multi-word expressions, difficulty is determined by the maximum SAMER level among their components. If the filtered list contains fewer than three valid words, the generation–validation cycle is repeated to ensure sufficient lexical constraints.

Vocabulary Relavance Fallback In cases where repeated validation still yields insufficient vocabulary items, we apply a semantic relevance model based on Sentence-BERT

[8] https://github.com/CAMeL-Lab/camel_tools, version 1.2.0

Table 3: Feature clusters used for evaluation

Cluster	Feature
Surface	total words
	avg words per sentence
	total unique words
	avg unique words
	avg word length
	max word length
	avg syllables per word
	max syllables
	avg unique lemmas
	avg unique lemma pos
	avg camel tokens
Dependency	IDF
	MOD
	OBJ
	SBJ
POS	ADJ
	ADP
	AUX
	CCONJ
	DET
	NOUN
	PUNCT
	VERB
DepPOS	IDF(NOUN)
	MOD(ADJ)
	MOD(ADP)
	MOD(AUX)
	MOD(CCONJ)
	MOD(DET)
	MOD(NOUN)
	MOD(PRON)
	MOD(PROPN)
	MOD(PUNCT)
	MOD(VERB)
	OBJ(ADJ)
	OBJ(ADV)
	OBJ(DET)
	OBJ(NOUN)
	SBJ(NOUN)
	SBJ(SCONJ)
	Tree Depth

(sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2) (Reimers and Gurevych, 2019). The model computes cosine similarity between the prompt topic and candidate SAMER lemmas within the appropriate readability level. Words exceeding a similarity threshold 0.50 and meeting minimum frequency constraints (frequency information is part of the SAMER Lexicon) are selected. These items are merged with previously validated words to produce the final vocabulary list used in constrained prompting.

5.2. Generation

All essays are generated using GPT-4o through the Chat Completions API in batch mode. Each batch entry corresponds to a single topic–CEFR pair. Each request includes a fixed system instruction in Arabic (أنت مساعد يكتب مقالات إنشائية عربية بمستويات قرائية مختلفة., *You are an assistant that writes Arabic essays at different readability levels.)* and the prompt. Generation is performed with a temperature of 0.7 to allow moderate linguistic variation while preserving control over readability constraints. All other parameters remain at default settings, and each prompt is generated once per prompting condition.

5.3. Profile Construction

The Dependency, POS and DepPOS cluster features listed in Table 3 are extracted using the CAMeL dependency parser (Elshabrawy et al., 2023). These features are computed through straightforward counting and averaging at the CEFR level, as described in Section 4. Similarly, surface-level length features are obtained using simple counts, such as word counts and length-based measures. To extract morphological and lexical features, we use the CAMeL Tools (Obeid et al., 2020) Modern Standard Arabic (MSA) BERT-based morphological disambiguation pipeline. This pipeline provides rich linguistic annotations, including diacritized forms, lemmas, part-of-speech tags, and morphological attributes. Lemma-based features are directly extracted from the disambiguator outputs. Tokenization is performed using CAMeL's `simple_word_tokenize` function to obtain the count of tokens. Syllable counts are computed by incorporating both phonological and morphological information. Specifically, we rely on the CAPHI (Consonant–Vowel pattern) representation (extracted from the disambiguation), the diacritized surface form, and morphological prefix annotations to obtain accurate syllable estimates. The CAPHI representation is segmented and examined for vowel units, with each vowel corresponding to a potential syllable. To improve accuracy, we apply the following linguistically motivated rules, following the approach of Rabih (2025):

- Final vowels corresponding to inflectional diacritics (حركات الإعراب) are excluded, as they reflect grammatical inflection rather than intrinsic word structure.
- Morphological prefixes such as the definite article (ال التعريف) and coordinating conjunction

Table 4: Cosine Similarity across dimensions.

Category	Group	Dimension / Level	P1	P2	P3	P4	P5
Overall	–	Overall	0.10	0.74	**0.91**	0.65	0.66
Cluster	–	Surface	0.21	0.76	**0.91**	0.83	0.86
	–	Dependency	0.11	0.77	**0.98**	0.63	0.89
	–	POS	-0.01	0.69	**0.90**	0.48	0.62
	–	DepPOS	0.08	0.73	**0.90**	0.61	0.56
Per Level	A	A1	-0.77	0.87	**0.95**	0.69	0.79
		A2		0.36	**0.90**		
	B	B1	0.76	-0.22	-0.23	-0.56	**0.59**
		B2		**0.96**	0.84	0.89	0.37
	C	C1	0.97	**0.97**	0.90	0.90	0.58
		C2		**0.97**	0.96	0.89	0.78

(واو العطف) are excluded, since they function as clitics and do not contribute to the core syllabic structure of the lexical stem.

- When CAPHI information is unavailable, syllables are estimated using a fallback heuristic that counts vowel-indicating diacritic characters in the diacritized word form.

6. Results and Analysis

We evaluate all five prompting conditions (P1–P5) using automatic readability prediction (BAREC), and feature-level deviation analysis against the reference CEFR-aligned profiles.

Table 4 reports cosine similarity between generated texts under the five prompting conditions (P1–P5) and the CEFR reference feature profiles. For robustness, evaluation is performed at the level of aggregated profiles rather than individual essays. Specifically, for each prompting condition and CEFR level, we compute the average feature vector across all generated essays for that level, and cosine similarity is then calculated between this averaged generated profile and the corresponding CEFR reference profile.

Overall, P3 achieves the highest alignment (0.91) and consistently outperforms other prompting strategies across feature clusters, particularly in the Dependency cluster (0.98), indicating strong syntactic conformity. At beginner levels (A1–A2), P3 shows the strongest alignment, while P2 performs competitively at advanced levels (C1–C2). Across feature categories, surface features show the most stable and consistently high alignment across prompting conditions, reflecting their direct responsiveness to explicit instructions. In contrast, syntactic features exhibit greater variability, while fine-grained DepPOS patterns remain the most challenging to control, as they emerge implicitly from writing style rather than being directly specifiable in prompts. However, the model appears to struggle at Level B1, where cosine similarities are comparatively low and even negative for several prompting conditions. This behavior can be explained by a systematic shift in the generated profiles toward higher complexity: instead of matching the reference B1 distribution, the model tends to produce texts with feature values closer to B2. As a result, many features that are below the reference mean at B1 become above the mean in the generated outputs, leading to opposite directional deviations after normalization and consequently negative cosine similarity. This indicates that the model overestimates intermediate-level complexity, effectively collapsing B1 toward the adjacent higher proficiency level.

Table 5: Average Taha-19 level cosine similarity per prompt.

Dimension	P1	P2	P3	P4	P5
Taha-19 level	0.26	0.83	**0.99**	0.93	0.91

Notably, prompting strategies that include explicit vocabulary additions tend to reduce overall performance compared to structure-focused prompting. This indicates that simply increasing lexical content does not necessarily improve CEFR alignment and may disrupt feature balance. One possible explanation is that under structural prompting, the model can generate text more naturally, loosely adhering to syntactic guidelines while maintaining fluent sentence construction. In contrast, when required to incorporate specific lexical items, the model tends to construct sentences around these words rather than generating text that organically conforms to the target syntactic profile. This effect is reflected in the sharp drop in POS cluster alignment under P4 (0.48) compared to P3 (0.90), as well as the decrease in DepPOS alignment from 0.90 to 0.61.

Table 5 reports the average cosine similarity be-

tween predicted BAREC levels and the reference CEFR levels across prompting conditions. Consistent with the feature-level analysis, P3 achieves the highest alignment (0.995), indicating near-perfect agreement with the target readability levels. P4 and P5 also show strong performance, while P2 performs moderately well. In contrast, P1 exhibits substantially lower alignment (0.2611). The large performance gap between P1 and P2 highlights the importance of explicit readability conditioning in LLM prompting. In P1, the model receives no signal about the target proficiency level, and therefore defaults to generating text at an average or internally preferred complexity level, which does not align with CEFR-specific linguistic profiles. These results further support the effectiveness of structured prompting in improving CEFR-level conformity.

A detailed table reporting the cosine similarity for each feature independently across prompting conditions is included in Appendix B.

7. Conclusions and Future Work

In this work, we presented the first multi-dimensional evaluation of CEFR-controlled Arabic text generation using LLMs. We proposed a comprehensive framework that combines automatic readability prediction (BAREC), lexical constraint validation, syntactic profiling, and feature-level cosine similarity analysis to assess whether instruction-following models can reliably generate CEFR-aligned Arabic texts. Our findings show that structured prompting substantially improves readability control. In particular, syntactically guided prompting (P3) consistently achieved the highest alignment with CEFR reference profiles across overall, cluster-level, and fine-grained BAREC evaluations. This study establishes an empirical foundation for integrating readability-aware generation into adaptive Arabic learning systems. By combining coarse CEFR categorization with fine-grained Taha-19 modeling, our framework enables scalable and interpretable evaluation of controlled text generation.

A promising direction for future research is extending this framework to the personalized learner level. Rather than relying solely on static CEFR-level profiles, individual student data could be used to construct learner-specific linguistic profiles, capturing strengths and weaknesses in vocabulary, syntax, and structural complexity. Generation prompts could then be dynamically adapted based on these personalized profiles, enabling fine-grained readability control tailored to individual learners. Such an approach is contingent upon the availability of sufficient and ethically collected learner data. Additionally, integrating readability-aware generation into real adaptive tutoring systems would enable longitudinal assessment of learning impact. In a "read–write–improve" loop, learners could receive automatically generated texts aligned to their evolving proficiency, produce written responses, and obtain feedback informed by the same profiling framework. Over time, this would allow continuous monitoring of linguistic development and systematic adjustment of generated materials. Evaluating such systems in real educational environments would provide deeper insight into how readability-controlled generation affects learning outcomes. Another direction for future work is evaluating multiple large language models. Since this study uses only GPT-4o, extending the analysis to other models would help assess the generalizability of the framework and determine whether the observed effects of structured prompting are consistent across different LLMs.

Ethics Statement

This work focuses on improving the reliability and safety of Arabic readability-controlled text generation for educational applications. All training and evaluation data used in this study are drawn from publicly available corpora (ZAEBUC, ARWI, SAMER, and BAREC), and no personally identifiable information was collected or processed.

Limitations

Despite providing the first multi-dimensional evaluation of CEFR-controlled Arabic text generation, this study has several limitations. First, although BAREC shows strong alignment with CEFR progression, our evaluation relies on a single readability model; future work could incorporate additional readability predictors to further validate and compare results.

Second, experiments are conducted using a single LLM (GPT-4o), which limits the generalizability of our findings. Future research should evaluate multiple LLMs to assess robustness across different architectures and decoding strategies.

8. Bibliographical References

Muhamed Al Khalil, Nizar Habash, and Zhengyang Jiang. 2020. A large-scale leveled readability lexicon for Standard Arabic. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3053–3062, Marseille, France. European Language Resources Association.

Violetta Cavalli-Sforza, Hind Saddiki, and Naoual Nassiri. 2018. Arabic readability research: Cur-

rent state and future directions. *Procedia Computer Science*, 142:38–49. Arabic Computational Linguistics.

C. o. E. Council of Europe. 2001. Common european framework of reference for languages: learning, teaching, assessment.

Mo El-Haj, Sultan Almujaiwel, Damith Premasiri, Tharindu Ranasinghe, and Ruslan Mitkov. 2024. Dares: Dataset for arabic readability estimation of school materials. In *Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context@ LREC-COLING 2024*, pages 103–113.

Khalid N. Elmadani, Nizar Habash, and Hanada Taha-Thomure. 2025. A large and balanced corpus for fine-grained Arabic readability assessment. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16376–16400, Vienna, Austria. Association for Computational Linguistics.

Ahmed Elshabrawy, Muhammed AbuOdeh, Go Inoue, and Nizar Habash. 2023. CamelParser2.0: A State-of-the-Art Dependency Parser for Arabic. In *Proceedings of The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*.

Nizar Habash and David Palfreyman. 2022. Zaebuc: An annotated arabic-english bilingual writer corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 79–88.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

Nouran Khallaf and Serge Sharoff. 2021. Automatic difficulty classification of arabic sentences. *arXiv preprint arXiv:2103.04386*.

Tarek Naous, Michael J Ryan, Anton Lavrouk, Mohit Chandra, and Wei Xu. 2024. Readme++: Benchmarking multilingual language models for multi-domain readability assessment. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2024, page 12230.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. CAMeL tools: An open source python toolkit for Arabic natural language processing. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France. European Language Resources Association.

Chatrine Qwaider, Bashar Alhafni, Kirill Chirkunov, Nizar Habash, and Ted Briscoe. 2025. Enhancing arabic automated essay scoring with synthetic data and error injection. *arXiv preprint arXiv:2503.17739*.

Nour Rabih. 2025. Noor at barec shared task 2025: A hybrid transformer-feature architecture for sentence-level readability assessment. In *Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks*, pages 331–342.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Leonardo FR Ribeiro, Mohit Bansal, and Markus Dreyer. 2023. Generating summaries with controllable readability levels. *arXiv preprint arXiv:2310.10623*.

Hanada Taha-Thomure. 2017. *Arabic Language Text Leveling* معايير هنادا طه لتصنيف مستويات النصوص(العربية). Educational Book House (دار الكتاب التربوي للنشر والتوزيع).

Sean Trott and Pamela Rivière. 2024. Measuring and modifying the readability of english texts with gpt-4. In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, pages 126–134.

Jeromie Whalen, Chrystalla Mouza, et al. 2023. Chatgpt: Challenges, opportunities, and implications for teacher education. *Contemporary Issues in Technology and Teacher Education*, 23(1):1–23.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text readability assessment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22, San Diego, CA. Association for Computational Linguistics.

A. CEFR Linguistic Profiles

Table 6: CEFR Linguistic Profiles

Feature	A1	A2	B1	B2	C1	C2
total_words	85.35	118.03	179.87	232.50	262.46	281.57
avg_words_per_sentence	6.06	10.29	15.69	17.48	18.23	18.59
total_unique_words	84.73	116.19	174.47	225.18	255.03	273.60
avg_unique_words	6.02	10.12	15.06	16.85	17.70	18.06
overall_avg_word_len	4.82	4.94	5.03	5.08	5.22	5.23
overall_max_word_len	9.04	9.51	10.26	10.58	11.06	11.10
avg_syllables_per_word	2.20	2.30	2.34	2.35	2.41	2.42
max_max_syllables	4.67	5.02	5.29	5.42	5.45	5.49
avg_unique_lemmas	5.97	9.93	14.62	16.44	17.37	17.74
avg_unique_lemma_pos	5.97	9.93	14.63	16.44	17.38	17.74
avg_camel_tokens	7.31	11.90	17.58	19.52	20.36	20.76
dep_IDF_mean	11.69	16.79	31.78	41.34	47.57	51.10
dep_MOD_mean	51.60	75.51	115.20	151.79	172.11	186.37
dep_OBJ_mean	32.99	48.82	64.41	80.54	86.22	92.24
dep_SBJ_mean	8.66	11.30	15.46	18.58	19.77	20.60
pos_ADJ_mean	11.03	15.29	25.59	35.15	43.17	47.64
pos_ADP_mean	12.19	21.29	34.56	44.82	49.29	51.86
pos_AUX_mean	0.49	0.85	1.01	1.31	1.62	1.68
pos_CCONJ_mean	6.78	12.63	18.18	23.59	25.38	28.25
pos_DET_mean	0.97	2.21	3.52	4.48	5.29	5.45
pos_NOUN_mean	38.17	52.92	88.26	114.68	130.50	139.61
pos_PUNCT_mean	17.87	18.74	22.98	28.13	30.79	33.08
pos_VERB_mean	18.39	22.42	23.04	27.55	29.13	31.26
depPOS_IDF(NOUN)_mean	5.98	9.97	22.55	31.04	37.81	40.72
depPOS_MOD(ADJ)_mean	6.92	10.12	20.26	28.49	36.23	40.07
depPOS_MOD(ADP)_mean	11.89	20.80	32.77	42.78	47.73	50.33
depPOS_MOD(AUX)_mean	0.49	0.85	1.01	1.31	1.62	1.68
depPOS_MOD(CCONJ)_mean	6.71	12.50	17.65	22.65	24.23	26.63
depPOS_MOD(DET)_mean	0.53	0.94	1.84	2.50	3.11	3.23
depPOS_MOD(NOUN)_mean	4.58	6.11	7.51	9.54	9.85	10.67
depPOS_MOD(PRON)_mean	0.59	1.23	2.44	3.64	4.09	4.31
depPOS_MOD(PROPN)_mean	0.04	0.08	0.18	0.27	0.35	0.35
depPOS_MOD(PUNCT)_mean	17.87	18.74	22.94	28.05	30.70	33.01
depPOS_MOD(VERB)_mean	0.79	2.07	5.62	8.77	10.57	12.04
depPOS_OBJ(ADJ)_mean	1.73	2.44	3.23	4.34	4.82	5.43
depPOS_OBJ(ADV)_mean	0.09	0.31	0.67	1.20	1.25	1.32
depPOS_OBJ(DET)_mean	0.07	0.48	0.91	1.26	1.37	1.38
depPOS_OBJ(NOUN)_mean	20.01	28.54	45.54	58.28	65.69	70.17
depPOS_SBJ(NOUN)_mean	5.23	6.14	10.31	13.23	14.94	15.66
depPOS_SBJ(SCONJ)_mean	0.40	0.56	1.96	2.03	2.20	2.28
tree_depth_mean	4.90	7.12	8.97	9.71	10.05	10.25
avg_ Taha-19_level	7.61	9.26	11.03	11.81	12.78	12.93

B. Feature-Level Evaluation

Table 7: Feature-Level Cosine Similarity per Prompt

Feature	P1	P2	P3	P4	P5
avg_camel_tokens	0.07	0.71	0.98	0.89	0.92
avg_level	0.26	0.83	1.00	0.93	0.91
avg_syllables_per_word	0.85	0.85	0.62	0.99	0.83
avg_unique_lemma_pos	0.08	0.71	0.99	0.88	0.91
avg_unique_lemmas	0.08	0.71	0.99	0.88	0.91
avg_unique_words	0.08	0.72	0.98	0.89	0.92
avg_words_per_sentence	0.08	0.72	0.98	0.89	0.93
dep_IDF_mean	0.17	0.81	0.99	0.73	0.98
dep_MOD_mean	0.05	0.77	0.99	0.70	0.89
dep_OBJ_mean	-0.02	0.63	0.98	0.46	0.91
dep_SBJ_mean	0.36	0.90	0.99	0.67	0.81
depPOS_IDF(NOUN)_mean	0.35	0.88	0.99	0.81	0.98
depPOS_MOD(ADJ)_mean	0.41	0.91	0.96	0.91	0.92
depPOS_MOD(ADP)_mean	0.06	0.74	0.99	0.63	0.94
depPOS_MOD(AUX)_mean	-0.13	0.60	0.51	-0.20	0.05
depPOS_MOD(CCONJ)_mean	0.03	0.72	0.94	0.74	0.47
depPOS_MOD(DET)_mean	-0.03	0.76	0.94	0.49	0.75
depPOS_MOD(NOUN)_mean	0.05	0.60	0.99	0.45	0.88
depPOS_MOD(PRON)_mean	-0.33	0.60	0.95	0.75	0.18
depPOS_MOD(PROPN)_mean	-0.28	0.87	0.96	0.66	0.63
depPOS_MOD(PUNCT)_mean	-0.09	0.70	0.88	0.45	0.39
depPOS_MOD(VERB)_mean	-0.07	0.68	0.97	0.67	0.63
depPOS_OBJ(ADJ)_mean	0.48	0.87	0.69	0.79	0.75
depPOS_OBJ(ADV)_mean	0.13	0.52	0.88	0.64	0.52
depPOS_OBJ(DET)_mean	0.93	0.83	0.94	0.92	0.59
depPOS_OBJ(NOUN)_mean	0.06	0.76	0.99	0.63	0.93
depPOS_SBJ(NOUN)_mean	0.47	0.94	0.95	0.78	0.82
depPOS_SBJ(SCONJ)_mean	1.00	0.95	0.95	0.92	0.65
depth_mean	-0.02	0.65	0.97	0.84	0.94
max_max_syllables	0.00	0.62	0.90	0.49	0.62
overall_avg_word_len	1.00	0.95	0.62	0.95	0.99
overall_max_word_len	0.53	0.92	0.99	0.94	0.97
pos_ADJ_mean	0.40	0.91	0.95	0.88	0.96
pos_ADP_mean	0.07	0.75	0.99	0.64	0.96
pos_AUX_mean	-0.13	0.60	0.51	-0.20	0.05
pos_CCONJ_mean	0.01	0.72	0.96	0.72	0.67
pos_DET_mean	0.14	0.76	0.96	0.68	0.97
pos_NOUN_mean	0.16	0.82	0.99	0.69	0.94
pos_PUNCT_mean	-0.08	0.70	0.88	0.46	0.40
pos_VERB_mean	-0.24	0.40	0.94	0.15	0.20
total_unique_words	0.11	0.79	0.99	0.66	0.92
total_words	0.11	0.79	0.99	0.67	0.93

C. CEFR Prompts and Vocabulary Lists

The data below can be useful for various educational applications [9].

Band	CEFR	Prompt	Vocabulary
Beginner	A	اوصف يومك المفضل Describe your favourite day	يوم, noun, day; some day; today صباح, noun, morning مساء, noun, evening شمس, noun, sun طعام, noun, food شراب, noun, drink صديق, noun, friend عائلة, noun, family وقت, noun, time نشاط, noun, activity بيت, noun, house …
Intermediate	B1	كيف يمكننا التعامل مع الضغوط النفسية How can we deal with psychological stress?	تعامل, noun, dealing استرخاء, noun, relaxation توتر, noun, tension قلق, noun, anxiety تغذية, noun, nutrition هواية, noun, hobby تحدي, noun, challenge …
	B2	كيف يمكننا التعامل مع الضغوط النفسية How can we deal with psychological stress?	التأمل, noun, contemplation الدعم, noun, support المرونة, noun, flexibility الوعي, noun, awareness التوازن, noun, balance التقييم, noun, evaluation التواصل, noun, continuity الاستشارة, noun, consultation الترفيه, noun, recreation التوجيه, noun, instruction …
Advanced	C1	تحدث عن أهمية التعليم الرقمي في عصرنا الحالي. Talk about the importance of digital education in our current era.	التعليم الرقمي, adj, digital التفاعل الرقمي, adj, digital المنصات التعليمية, adj, educational الكفاءة التقنية, noun, technology الأدوات الرقمية, adj, digital المهارات الرقمية, adj, digital تطوير الذات, noun, self-development التعليم التفاعلي, adj, interactive العصر الرقمي, adj, digital era التعلم الذاتي, adj, autonomous …
	C2	تحدث عن أهمية التعليم الرقمي في عصرنا الحالي. Talk about the importance of digital education in our current era.	الأمان السيبراني, adj, -- التعلم مدى الحياة, noun, -- المحاكاة, noun, imitation الواقع المعزز, adj, augmented التعليم المدمج, adj, blended …

Figure 5: Example of CEFR-aligned essay prompts with vocabulary lists.

[9] https://github.com/noorrabih/CEFR-Controlled-Arabic-Generation-Data.git

D. Prompts

ID	Arabic Prompt	English Prompt
P1	اكتب مقالًا إنشائيًا عربيًا متكاملًا. يجب أن يتناول المقال الموضوع التالي: {topic_text}. يجب أن يكون النص مترابطًا وسلسًا في فقرة أو عدة فقرات متتابعة، ويعكس بنية منطقية تبدأ بتمهيد للفكرة ثم تطويرها ثم إنهائها بخلاصة طبيعية، من دون استخدام عناوين أو كلمات دالة مثل (مقدمة، محتوى، خاتمة) أو أي تقسيم صريح.	Write a complete Arabic composition essay The essay must address the following topic: {topic_text}. The text should be coherent and fluent in one or more connected paragraphs, and it should follow a logical structure that begins by introducing the idea, then developing it, and finally ending with a natural conclusion, without using headings or explicit markers such as (Introduction, Body, Conclusion) or any explicit sectioning.
P2	اكتب مقالًا إنشائيًا عربيًا متكاملًا بمستوى قابلية قراءة {level}. يجب أن يتناول المقال الموضوع التالي: {topic_text}. يجب أن يكون النص مترابطًا وسلسًا في فقرة أو عدة فقرات متتابعة، ويعكس بنية منطقية تبدأ بتمهيد للفكرة ثم تطويرها ثم إنهائها بخلاصة طبيعية، من دون استخدام عناوين أو كلمات دالة مثل (مقدمة، محتوى، خاتمة) أو أي تقسيم صريح.	Write a complete Arabic composition essay at readability level {level}. The essay must address the following topic: {topic_text}. The text should be coherent and fluent in one or more connected paragraphs, and it should follow a logical structure that begins by introducing the idea, then developing it, and finally ending with a natural conclusion, without using headings or explicit markers such as (Introduction, Body, Conclusion) or any explicit sectioning.
P3	اكتب مقالًا إنشائيًا متكاملًا بمستوى قابلية قراءة {level}. يجب أن يكون المقال عن الموضوع التالي: {topic_text}. حاول أن يكون النص قريبًا من الخصائص العامة التالية: الطول والبنية: ● طول النص قريب من {total_words}. ● طول الجمل قريب من {avg_words_per_sentence}. ● عدد الكلمات الفريدة قريب من {total_unique_words}. ● متوسط طول الكلمة (عدد الحروف): {overall_avg_word_len}. الوصف والتفصيل: ● عدد الصفات قريب من {pos_ADJ_mean}. ● استخدم صفات لوصف الأشخاص والأشياء والأفكار. ● كلما زاد المستوى، زادت التفاصيل. بناء الجملة والمعنى: ● عدد الفاعل قريب من {dep_SBJ_mean}. ● عدد المفاعيل والأمثلة قريب من {dep_OBJ_mean}. ● في المستويات المتقدمة (B1-C2)، اربط السبب بالنتيجة. البنية النحوية: ● العمق الأدنى لشجرة التحليل قريب من {max_depth_low}. ● العمق الأقصى لشجرة التحليل قريب من {max_depth_high}. ● كلما ارتفع المستوى، استخدم جملًا ذات بنية أعمق وأكثر تفرعًا. الترابط والمنطق: ● استخدم روابط منطقية تناسب المستوى. ● في المستويات المتقدمة، استخدم التعليل والمقارنة. الأسلوب العام: ● لا تستخدم الأسلوب الحواري إلا في B2 أو أعلى. ● تجنّب التعجب والمبالغة في المستويات الدنيا. ● يجب أن يعكس الأسلوب مستوى متعلم حقيقي. قيود مهمة:	Write a complete composition essay at readability level {level}. The essay must be about the following topic: {topic_text}. Try to make the text close to the following general characteristics: Length and structure: ● Total length should be close to {total_words}. ● Sentence length should be close to {avg_words_per_sentence}. ● Number of unique words should be close to {total_unique_words}. ● Average word length (number of characters): {overall_avg_word_len}. Description and detail: ● Number of adjectives should be close to {pos_ADJ_mean}. ● Use adjectives to describe people, objects, and ideas. ● As the level increases, include more details. Sentence construction and meaning: ● Number of subjects should be close to {dep_SBJ_mean}. ● Number of objects and examples should be close to {dep_OBJ_mean}. ● At advanced levels, link cause and effect. Syntactic structure (parse tree): ● Minimum parse tree depth should be close to {max_depth_low}. ● Maximum parse tree depth should be close to {max_depth_high}. ● As the level increases, use deeper and more branched sentence structures. Cohesion and logic: ● Use logical connectors appropriate for the level. ● At advanced levels, use justification and comparison. General style: ● Do not use a conversational style unless the level is B2 or higher. ● Avoid exclamation and exaggeration at lower levels. ● The style should reflect a real learner at the specified level. Important constraints: ● These values are guidelines only.

ID	Arabic Prompt	English Prompt
	• هذه القيم إرشادية فقط. • لا تذكر أي أرقام في النص. • لا تشرح هذه القيود في الناتج. هيكل النص: • مقدمة ومحتوى وخاتمة بدون عناوين صريحة. • فقرات متصلة فقط.	• Do not mention any numbers in the generated text. • Do not explain these constraints in the output. Text structure: • Introduction, body, and conclusion without headings. • Connected paragraphs only.
P4	اكتب مقالًا إنشائيًا متكاملًا بمستوى قابلية قراءة {level}. يجب أن يكون المقال عن الموضوع التالي: {topic_text}. تأكد من إدخال بعض الكلمات المفتاحية التالية بشكل طبيعي داخل النص لإنشاء قصة متماسكة جميلة: {words}. يجب أن يتكوّن المقال من مقدمة ومحتوى وخاتمة دون أي عناوين أو تقسيمات ظاهرة. اكتب النص كسرد متصل في فقرات متتابعة فقط دون أي تنسيق أو عناوين فرعية. احرص على أن يكون المقال مترابطًا ومنظمًا ومناسبًا لمستوى القراءة المطلوب. تأكد من أن جميع الجمل منطقية ومترابطة من حيث المعنى. تجنّب أي عبارات أو أفكار غير واقعية أو غير منطقية. تأكد من أن الأمثلة والحقائق تتوافق مع العادات والثقافة والمنطق العام. احرص على أن يكون النص واضحًا وسهل الفهم دون أخطاء لغوية أو تركيبية.	Write a complete composition essay at readability level {level}. The essay must be about the following topic: {topic_text}. Naturally incorporate the following keywords into the text to create a coherent and engaging narrative: {words}. The essay must consist of an introduction, body, and conclusion without any headings or visible section divisions. Write the text as continuous connected paragraphs only, without formatting or subheadings. Ensure the essay is coherent, well-organized, and appropriate for the requested readability level. Make sure all sentences are logical and connected in meaning. Avoid unrealistic or illogical statements or ideas. Ensure examples and facts are consistent with cultural norms and general common sense. Ensure the text is clear, easy to understand, and free of grammatical or structural errors.
P5	اكتب مقالًا إنشائيًا متكاملًا بمستوى قابلية قراءة {level}. يجب أن يكون المقال عن الموضوع التالي: {topic_text}. تأكد من إدخال بعض الكلمات المفتاحية التالية بشكل طبيعي داخل النص لإنشاء قصة متماسكة جميلة: {words}. حاول أن يكون النص قريبًا من الخصائص العامة التالية: Specifications from p3, omitted due to space limitations	Write a complete composition essay at readability level {level}. The essay must be about the following topic: {topic_text}. Naturally incorporate the following keywords into the text to create a coherent and engaging narrative: {words}. Try to make the text close to the following general characteristics: Specifications from p3, omitted due to space limitations

Figure 6: Prompts used for controlled Arabic text generation at different readability levels.

Lexical Conditioning of Model's Distribution through Uncertainty-gated Soft-Mixing of Probabilities

Michele Papucci[1,2], **Giulia Venturi**[2], **Felice Dell'Orletta**[2]
[1]University of Pisa
[2]ItaliaNLP @ Institute for Computational Linguistics "A. Zampolli" (CNR-ILC), Pisa
michele.papucci@phd.unipi.it, {giulia.venturi, felice.dellorletta}@ilc.cnr.it

Abstract

We present Uncertainty-Gated Lexical Decoding (UGLD), a decoding-time framework for fine-grained lexical control in Large Language Models (LLMs) that explicitly addresses the trade-off between controllability and fluency. UGLD adaptively scales intervention through an entropy-based gating mechanism derived from the model's predictive distribution, activating control when uncertainty is high and limiting interference when predictions are confident. The method supports both promotion toward and against predefined vocabularies. We evaluate UGLD in Italian on two open-weight LLMs (ANITA 8B and Qwen 3 4B) across paraphrasing and free-text generation settings, considering Simple Vocabulary Conditioning and Jargon Reduction scenarios. Automatic evaluation shows consistent improvements in lexical coverage over standard decoding strategies, while human evaluation confirms that fluency is preserved under controlled intervention.

Keywords: Controlled Text Generation, Lexically Constrained Decoding, Entropy-Gated Decoding

1. Introduction

Controlling the behavior of Large Language Models (LLMs) for Text Generation is becoming increasingly important as these models are used in a growing number of real-world scenarios that require adherence to multiple constraints. To address this need, recent techniques for Controlled Text Generation (CTG) include fine-tuning approaches (Nguyen et al., 2024), prompt-based methods that express the constraint as a natural-language instruction (Zhou et al., 2023), and weighted decoding strategies that intervene directly in the model's output distribution to adjust token probabilities during the decoding stage (Pascual et al., 2021; Yang and Klein, 2021). While the latter approaches are particularly suitable, especially in low-resource settings, they face the recurrent challenge of maintaining fluency of the generated texts. In fact, prior work has shown that as the control strength increases beyond a certain threshold, fluency can rapidly decrease (Zhong et al., 2023).

Building on these premises, we introduce *Uncertainty-Gated Lexical Decoding*, a decoding-time approach for fine-grained lexical control that aims to preserve the fluency of the generated text. The core idea is to modulate the strength of decoding interventions according to the model's uncertainty in next-token prediction, rather than applying lexical constraints uniformly across the generation process. To this end, we devise an *uncertainty-gated* mechanism derived from the entropy of the model's predictive distribution, which activates control primarily when the model is uncertain and limits interference when it is confident in its predictions. The gate supports two complementary forms of lexical control: conditioning generation *towards* a predefined set of lexical items through explicit lexical priors, and conditioning it *against* undesired vocabulary through logit-level penalties that suppress specific words.

We evaluate the proposed approach in two CTG settings: **paraphrasing**, where the model rewrites an input sentence under lexical constraints, and **free-text generation**, where constraints are applied during unconstrained continuation. Experiments are conducted on Italian and focus on two complementary forms of lexical conditioning that are relevant in real-world applications: **Simple Vocabulary Conditioning**, which involves guiding generation towards a predefined lexicon of simple and high-accessibility words, and **Jargon Reduction**, which aims to guide generation away from domain-specific terminology by suppressing technical lexical items. We consider these two conditioning scenarios because they represent complementary directions of lexical control. Together, they can be viewed as building blocks that could be integrated into a broader Controlled Text Simplification framework, which we leave as future work.

Contributions: *i)* we propose a novel decoding-time framework that adaptively modulates lexical control through an entropy-based gating mechanism; *ii)* we make use of the lexical control mechanism to condition two LLMs towards and against a predefined vocabulary, without requiring additional training or external discriminators, making it particularly suitable for low-resource settings and

Proceedings of the Joint Workshop on Readability and Text Simplification (READIxTSAR) @ LREC 2026, pages 89–100
11 May 2026.

languages[1]; *iii)* we provide a twofold evaluation consisting of an automatic coverage metric to quantify lexical shifts across different levels of intervention strength and human judgments to assess fluency preservation.

2. Related Works

Decoding techniques have been used as a way to control LLMs' generation, and are widely considered a family of Controlled Text Generation (CTG) techniques (Zhang et al., 2023). Relevant previous works fall broadly into hard-constraint or soft-constraint decoding methods.

Hard constraint decoding methods, such as Grid Beam Search and Dynamic Beam Allocation (Hokamp and Liu, 2017; Post and Vilar, 2018), enforce the presence of specific words or phrases through structural modifications of the beam search. While effective at guaranteeing constraint satisfaction, they do not provide adaptive control, leading to less fluent text.

Soft-constraint approaches, including GeDI, PPLM, FUDGE, and DExperts (Krause et al., 2021; Pascual et al., 2021; Yang and Klein, 2021; Liu et al., 2021), modify token probability to induce high-level attributes (e.g., sentiment, toxicity, topic). These methods, however, rely on a fixed strength intervention and/or on a 'learned' discriminator to manipulate the model's logits. Such discriminators are often implemented as trained classifiers, which require additional annotated data and training resources that may not be available in low-resource settings, or as auxiliary LLMs, which substantially increase computational cost at inference time. Our method instead uses *explicit lexical priors* and *logit penalties* with an *adaptive intervention strength modulated by the model's own uncertainty*, and does not require training or querying external models. More generally, incorporating uncertainty into decoding decisions has been explored in prior work through entropy-based measures, which provide a model-internal signal of confidence during language generation, although these approaches are not typically formulated within a CTG setting. Locally Typical Sampling (Meister et al., 2023) restricts the sampling to tokens whose information content matches the model's conditional entropy to mimic the human information pacing. Similarly, entropy-based methods such as EGED (Das et al., 2025) adjust decoding behavior under uncertainty, but do not attempt lexical or attribute steering. ECO decoding (Shin et al., 2025) is closest to our method: it adapts attribute-control strength using entropy, but still assumes a classifier-based attribute model rather than explicit vocabularies.

3. Methodology

We present a decoding technique for Controlled Text Generation called **Uncertainty-Gated Lexical Decoding** (UGLD). In particular, we present two variations of UGLD: one that conditions the model's distribution towards producing a pre-set vocabulary of tokens (UGLD-t), the other conditions the model's distribution against the chosen vocabulary (UGLD-a). To do that, while preserving the model's fluency, we scale the intervention strength by the model's uncertainty in the next-token prediction, leading to strong conditioning only when the model is uncertain. To model uncertainty, we employ Shannon's Entropy $H(\mathbf{p})$:

$$H(\mathbf{p}) = -\sum_i p_i \log p_i$$

where $\mathbf{p}$ is a probability vector of the size of the model's vocabulary $\mathcal{V}$, obtained by soft-maxing the last layer model logits $\mathbf{z}$, i.e., $\mathbf{p} = \text{SoftMax}(\mathbf{z})$ before applying any probability manipulations.
By using the entropy $H(\mathbf{p})$ we define a smooth gating mechanism $\phi(\mathbf{p}) \in [0, 1]$, that determines how strongly to condition the distribution:

$$\phi(\mathbf{p}) = \sigma(\frac{H(\mathbf{p}) - \tau}{s}) \quad (1)$$

Here, τ acts as an entropy threshold when the intervention is activated. For $H(\mathbf{p}) \ll \tau$, the gate is closed $\phi \approx 0$; for $H(\mathbf{p}) \gg \tau$ the gate is open $\phi \approx 1$. Since $H(\mathbf{p}) \in [0, \log|\mathcal{V}|]$, τ is typically chosen in $[0, \log|\mathcal{V}|]$, as outside this range we have degenerate behaviors[2]. Finally, σ is the sigmoid function that normalizes the gate $\phi(\mathbf{p}) \in [0, 1]$ and s is a strictly positive smoothing factor that controls *how fast* the gate opens: a smaller s value makes the sigmoid reach 1 with lower values of $H(\mathbf{p})$.

Conditioning Towards (UGLD-t) Given a subset of tokens that we want the LLM to generate with **higher** relative probability than the rest of the vocabulary, hereafter referred to as *green tokens* ($\mathcal{V}_{green} \subset \mathcal{V}$), we employ a soft mixture of probability distributions to shift the model's output distribution toward the green tokens when the uncertainty is high:

$$\mathbf{p}' = (1 - \alpha)\mathbf{p} + \alpha\mathbf{q} \quad (2)$$

Here $\mathbf{p}$ and $\mathbf{q}$ are the two probability distributions we are mixing. In particular, $\mathbf{p}$ is the probability

[1] UGLD is available on GitHub: https://github.com/michelepapucci/ugld and can be installed with PyPI: https://pypi.org/project/ugld/

[2] When $\tau \gg \log|\mathcal{V}|$ we approximate an always closed gate, and with $\tau \ll 0$ we approximate an always open gate

vector over the vocabulary produced at the current decoding step, while $\mathbf{q}$ is our conditioning prior distribution that allocates all of its probability mass to the *green* tokens. Formally, $\mathbf{q}$ is any valid probability distribution where $q_i = 0 \; \forall i \notin \mathcal{V}_{green}$ and $\sum_i q_i = 1$. Finally, α represents the intervention strength, calculated as the maximum intervention strength allowed $\alpha_{max} \in [0, 1]$ weighted by the entropy gate $\phi(\mathbf{p})$:

$$\alpha = \phi(\mathbf{p})\alpha_{max}$$

Because $\alpha \in [0, 1]$, this update forms a convex combination of the two distributions, ensuring that $\mathbf{p}'$ remains a valid probability distribution. Possible instantiations of the conditioning distributions $\mathbf{q}$ are described in Section 4.2.

Conditioning Against (UGLD-a) Given a subset of tokens that we want the LLM to generate with **lower** relative probability than the rest of the vocabulary, hereafter referred to as *red tokens* ($\mathcal{V}_{red} \subset \mathcal{V}$), we manipulate the model-produced logits to penalize the generation of those tokens when the uncertainty is high. To avoid normalization artifacts or negative probabilities, the negative conditioning is performed directly in **logits space**. Given the model's logits $\mathbf{z}$, we can subtract a vector of penalties $\lambda\mathbf{r}$:

$$\mathbf{z}' = \mathbf{z} - \lambda\mathbf{r}$$

where $\mathbf{r}$ is a vector of size $|\mathcal{V}|$ that contains weights for each token inside the *red* vocabulary. Specifically, $r_i > 0 \; \forall i \in \mathcal{V}_{red}$ and $r_i = 0$ otherwise. λ is instead the penalty strength that gets scaled for each token by $\mathbf{r}$, and is calculated as the maximum penalty strength allowed $\lambda_{max} \geq 0$ weighted by the entropy gate $\phi(\mathbf{p})$:

$$\lambda = \phi(\mathbf{p})\lambda_{max}$$

The newly obtained logits $\mathbf{z}'$ can be SoftMaxed to obtain the final probability distribution over the vocabulary from which the next-token prediction can be sampled. Possible ways to build the weight vector $\mathbf{r}$ are described in Section 4.2.

4. Experimental Settings

The effectiveness of the Uncertainty-Gated Lexical Decoding strategy was evaluated in two conditioning scenarios. The Conditioning Towards strategy was tested in a **Simple Vocabulary Conditioning** scenario, which evaluates the ability of the strategy to guide the model's generation towards a predefined set of simple and accessible words, while the Conditioning Against strategy was tested in a **Jargon Reduction** scenario, which evaluated the ability to guide generation away from domain-specific terminology by suppressing technical vocabulary. In both scenarios, the decoding strategy was tested in two controlled text generation settings: **paraphrasing**, where an original sentence is rewritten under lexical conditioning, and **free-text generation**, where lexical conditioning is applied while generating a short paragraph as a continuation of an input sentence[3].

All experiments were conducted using two open-weight LLMs: one fine-tuned for Italian and one multilingual. Specifically, we considered ANITA 8B (Polignano et al., 2024)[4], a model fine-tuned on Italian data and based on LLaMA 3 (Grattafiori et al., 2024), and Qwen 3 4B (Yang et al., 2025)[5], a multilingual LLM pre-trained on large-scale multilingual corpora, supporting Italian generation and used here in its instruction-tuned version.

4.1. Data and Lexical Resources

Simple Vocabulary Conditioning. We considered a set of 100 sentences randomly sampled from Wikipedia page dumps. We selected Wikipedia as our source data because it provides a heterogeneous collection of general-purpose texts. The two tested LLMs were conditioned toward generating *green tokens* from the *New Basic Italian Vocabulary* (NBIV) (De Mauro and Chiari, 2016), designed to include the vocabulary that ensures high readability for readers with heterogeneous education levels and originally including a list of 7,000 words highly familiar to native speakers of Italian. Specifically, we considered the subset of words belonging to the class of *Fundamental Words*, which includes very frequent and widely used words in Italian. Within this class, we included in the list of conditioning green tokens only words categorized as adjectives, nouns, verbs, and past and present participles. This choice is motivated by the intuition that content words, rather than grammatical ones, carry the core semantic load of a sentence and therefore represent the most effective targets for lexical conditioning aimed at increasing accessibility. To convert this list from a list of lemmas to a list of tokens, we first expanded each lemma to every possible form, obtaining a final list of $98\,430$ word forms, then converted it to token IDs using the tokenizer for the selected model, obtaining 4686 tokens for Anita and 4654 tokens for Qwen.

Jargon Reduction. Although the scenario is in principle domain-general, we adopt the medical do-

[3] This was achieved through prompting. For the paraphrasis, we asked the model to rewrite the sentence in its own words. while for free-text generation we asked the model to continue the sentence with a short paragraph.

[4] swap-uniba/LLaMAntino-3-ANITA-8B-Inst-DPO-ITA

[5] Qwen/Qwen3-4B-Instruct-2507

main as representative, as it has been widely considered in studies aimed at reducing the complexity of expert-specific jargon. To this end, we used the Italian version of the manuals available on MSD Manuals[6], one of the world's most widely used specialized websites offering publicly accessible medical information for both healthcare professionals (Professional version) and non-expert users (Consumer version), written to be clear and accessible to non-expert readers. We considered a total of 100 sentences. As a lexical resource, we consider the lexicon deriving from the Professional and Consumer versions of the MSD articles. Specifically, we employed a contrastive TF-IDF variant. We computed Term Frequency (TF) on the professionals' corpus by aggregating and length-normalizing term counts across all professional documents. We then computed Inverse Document Frequency (IDF) on the consumers' corpus, measuring how many consumer documents each term appears in (with smoothing). Candidate terms were extracted from professionals' texts and ranked by multiplying their professional TF by their consumer-based IDF. Intuitively, this procedure extracts words that are frequent in professionals' texts but rare across consumers' documents. In practice, it highlights terminology that is characteristic of the professional domain, such as technical, clinical, or specialized expressions, while downweighting words that are common in both groups. Finally, we expanded each word to all possible forms of its lemma, obtaining $111\,196$ forms, and extracted the token IDs using the tokenizer for the selected model, yielding $11\,820$ tokens for Anita and $11\,727$ for Qwen.

4.2. UGLD Hyperparameter Choice

To test UGLD in both scenarios, we had to choose a number of hyperparameters.

Simple Vocabulary Conditioning. For this scenario, we employed UGLD-t, to condition the LLMs towards generating more tokens that appear in the NBIV. For the choice of conditioning prior $\mathbf{q}$ (See Eq. 2), we experimented with three possible choices:

1. **Uniform** We build a uniform prior that spreads the probability mass equally across all green tokens in $\mathcal{V}_{green}$:

$$q_i = \begin{cases} \frac{1}{|\mathcal{V}_{green}|} & i \in \mathcal{V}_{green} \\ 0, & \text{otherwise} \end{cases}$$

2. **Top-K** We create a uniform prior that only spreads its probability mass across the top-K candidates in V_{green} in terms of probability of being produced at that time step:

$$q_i = \begin{cases} \frac{1}{|K|} & i \in \text{Top-}K(\mathcal{V}_{green}) \\ 0, & \text{otherwise} \end{cases}$$

3. **Re-normalizing** We re-normalize the model probability distribution at each time step over the set of green tokens V_{green}:

$$q_i = \begin{cases} \frac{p_i}{\sum_{j \in \mathcal{V}_{green}} p_j} & i \in \mathcal{V}_{green} \\ 0, & \text{otherwise} \end{cases}$$

We then set the threshold τ based on the distribution of entropy values $H(\mathbf{p})$ in the dataset (See Figure 1a) so that the entropy is higher than τ in around a third of the decoding steps. We chose 0.1 for Qwen, and 0.5 for Anita. Then, for α_{max} and s we tested each prior in three different scenarios:

- **Soft Conditioning**. We kept the intervention soft, with a max intervention strength $\alpha_{max} = 0.25$ and a relatively fast smoothing $s = 0.3$;
- **Medium Conditioning**. Here we increased the max intervention strength $\alpha_{max} = 0.5$ and kept the smoothing $s = 0.3$;
- **Strong Conditioning**. Finally, we pushed the intervention strength to $\alpha_{max} = 0.8$ and made the decoding transition very rapidly towards the maximum strength of intervention with a very fast smoothing $s = 1^{-12}$;

Jargon Reduction. For the task, we employed UGLD-a to condition the model to generate fewer technical words. To do that, we experimented with two different ways of constructing the weight vector $\mathbf{r}$:

1. **Fixed Weights**: A fixed penalty weight of 1 is applied to each token in V_{red}:

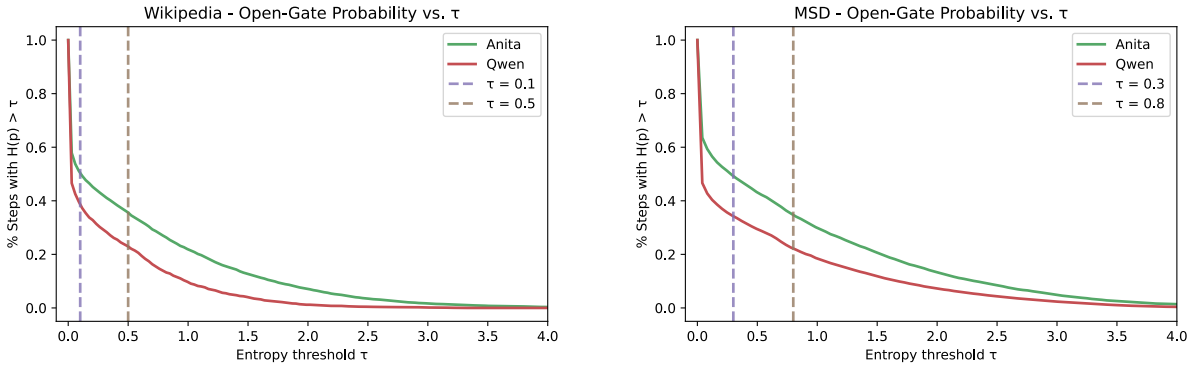
$$r_i = \begin{cases} 1 & \text{if } i \in R \\ 0 & \text{otherwise} \end{cases}$$

2. **Dynamic Weights**: A dynamic penalty is calculated at each time step, where a larger penalty is allocated towards red tokens that the model is currently more likely to generate. For each $i \in |\mathcal{V}_{red}|$ we compute a relative score via min–max normalization over the probabilities of red tokens p^{red} and map it to $[1, 2]$:

$$r_i = \begin{cases} 1 + \frac{p_i - \min(p^{red})}{\max(p^{red}) - \min(p^{red})} & \text{if } i \in \mathcal{V}_{red} \\ 0 & \text{otherwise} \end{cases}$$

Similar to the Simple Vocabulary Conditioning scenario, we fixed τ to 0.3 for Qwen and 0.8 for

[6] https://www.msdmanuals.com



(a) Wikipedia

(b) MSD

Figure 1: Percentage of decoding steps with entropy higher than possible τ threshold values. On the left, the entropy calculated on Wikipedia texts, on the right, the entropy calculated on MSD texts.

Anita based on the model's entropy distribution over the decoding steps of the dataset (see Figure 1b). Then, for λ_{max} and s we created again three scenarios:

- **Soft Conditioning**. We tried a soft penalty to the logits, with a $\lambda_{max} = 2$ and a relatively fast smoothing $s = 0.3$;
- **Medium Conditioning**. Here we increased the max penalty $\lambda_{max} = 4$ and kept the smoothing $s = 0.3$;
- **Strong Conditioning**. Finally, we pushed the penalty $\lambda_{max} = 6$ and made the decoding transition very rapidly towards the maximum penalty, with a very fast smoothing $s = 1^{-12}$;

4.3. Evaluation Methods

We adopted two types of evaluation methods, i.e., automatic metrics and human judgments, aimed at assessing complementary aspects of the UGLD approach. Automatic metrics are considered to quantify the extent to which the approach successfully guides the model's lexical choices toward or away from predefined vocabularies under different decoding settings. On the contrary, human judgments are aimed at assessing fluency of the generated text, given that decoding-time control may negatively affect this aspect of the generation process.

Automatic evaluation. For both controlled generation settings (paraphrasing and free-text generation), we computed the coverage of either green or red tokens in the generated outputs under different conditioned decoding strategies, considering both greedy decoding and nucleus sampling. Specifically, The coverage is calculated as the length-normalized percentage of generated tokens that are part of the selected lexicon, so for the the Simple Vocabulary Conditioning it is the proportion of generated green tokens belonging to the target accessible lexicon[7] while for Jargon Reduction, we measured the proportion of generated red tokens belonging to the technical vocabulary. For each setting, we compared the baseline decoding behavior (Greedy and Nucleus) with its uncertainty-gated counterpart (+ UGLD-t or + UGLD-a), ensuring that observed differences can be attributed to the proposed conditioning strategy rather than to decoding variability.

Human judgments. While automatic evaluation was conducted on the full set of experimental configurations, human judgments were collected on a selected subset. We focused on the Simple Vocabulary Conditioning scenario, as it involves paraphrasing and free-text generation of Wikipedia-style sentences, which are more easily assessable by human evaluators than texts from specialized medical domains. Human evaluation was further restricted to comparing the Nucleus baseline with Nucleus+UGLD-t under the Re-normalizing configuration. This choice was made as the Nucleus baseline is a very commonly used decoding strategy used in the wild. As for the other hyperparameters, we considered only the Soft and Strong conditioning settings, corresponding to the two extremes of conditioning strength, in both paraphrasing and free-text scenarios. This resulted in 50 pairs per model (Qwen and ANITA) per scenario, for a total of 400 pairs. For each model and scenario, the pairs were randomly distributed into 2 questionnaires of 25 pairs each, administered to 5 distinct annotators recruited via Prolific, all native speakers of Italian, for a total of 80 annotators[8]. For each pair,

[7] Table 6 (in the Appendix) shows an example in paraphrasing a sentence using different models and different UGLD-t configurations.

[8] Annotators were compensated £7.50/h. For the paraphrasis setting, we estimated a completion time of 20 min. while the observed average completion time was

Green Token Coverage on Wikipedia (Paraphrasis) ↑

Original Text		0.37								
Anita	**Greedy**	0.38								
	Nucleus	0.37								
		Uniform			**Top-K**			**Re-Normalization**		
		Soft	Medium	Strong	Soft	Medium	Strong	Soft	Medium	Strong
	Greedy + UGLD-t	0.38	0.38	0.38	0.38	0.38	0.40	0.38	0.38	**0.44**
	Nucleus + UGLD-t	0.37	0.37	0.36	0.40	0.40	0.40	0.39	0.39	0.40
Qwen	**Greedy**	0.38								
	Nucleus	0.37								
		Uniform			**Top-K**			**Re-Normalization**		
		Soft	Medium	Strong	Soft	Medium	Strong	Soft	Medium	Strong
	Greedy + UGLD-t	0.38	0.38	0.38	0.38	0.38	0.40	0.39	0.39	**0.42**
	Nucleus + UGLD-t	0.37	0.38	0.37	0.37	0.38	0.37	0.39	0.38	0.41

Table 1: In gray the best greedy configuration, in blue is the best nucleus configuration. In **bold** the best configuration overall, per-model.

annotators answered 2 binary (yes/no) questions assessing the fluency of each sentence in terms of grammatical correctness ("Is sentence 1 grammatically correct?" / "Is sentence 2 grammatically correct?"). The UGLD and Nucleus sampled sentences came from the same prompt, model, and scenario, with only the decoding strategy changed, and were randomly assigned to be either sentence 1 or sentence 2.

5. Results

Automatic evaluation. Results are reported in Tables 1, 2, 3, and 4. Tables 1 and 2 present results for the paraphrasing setting, reporting lexical coverage in the conditioning-toward scenario (measured as green-token coverage) and the conditioning-against scenario (measured as red-token coverage) for Wikipedia and MSD texts, respectively. Tables 3 and 4 report the corresponding results for free-text generation. For all settings, we consider greedy decoding and nucleus sampling as baselines. The latter coverage was obtained as the average coverage across 9 generation runs. We then applied UGLD-t to Wikipedia and UGLD-a to MSD, and after manipulating the probability distribution with the two techniques, we decoded both by greedy decoding (Greedy + UGLD) and nucleus decoding (Nucleus + UGLD)[9].

As a general outcome, we can observe that **free-text generation exhibits stronger lexical control effects than paraphrasing**. Specifically, free-text generation yields higher green token coverage in the conditioning-toward scenario (UGLD-t) and lower red token coverage in the conditioning-against scenario (UGLD-a). This pattern is consistent with the entropy-based design of UGLD: free-text generation typically involves higher uncertainty, which increases the activation of the gating function and results in stronger lexical control. In contrast, paraphrasing constrains the model through the source sentence, reducing uncertainty and therefore limiting the effect of the conditioning.

More specifically, we can observe a number of differences across conditioning scenarios. In the conditioning-toward scenario (Wikipedia), baseline green-token coverage in paraphrasing remains close to that of the original sentence for both ANITA and Qwen. Under UGLD-t, modest gains are observed with Uniform and Top-K priors, while Re-Normalization yields more substantial improvements, particularly when combined with greedy decoding. A similar trend emerges in free-text generation, where stronger hyperparameter settings further increase green-token coverage for both models. In contrast, in the second conditioning scenario (MSD), conditioning against the red vocabulary (UGLD-a) produces markedly larger shifts relative to baseline, for both models. Although Nucleus paraphrasing already reduces red-token coverage compared to the original text, UGLD-a further decreases it for both models, especially when combined with Dynamic Weights under Strong hyperparameter settings. In free-text generation, this effect becomes even more pronounced. Overall, these results indicate that **conditioning against a vocabulary produces larger coverage differences** than conditioning toward a vocabulary, across both generation settings. It seems to suggest that suppression tends to induce larger shifts because directly penalizing selected tokens forces a stronger redistribution of token probabilities, whereas promotion must compete with the model's original token ranking, resulting in more moderate deviations from

21 minutes and 24 sec. For the free-text setting, we estimated a completion time of 30 min. while the observed average completion time was 32 minutes and 39 sec.

[9]We chose a top-p = 0.9, which is a common choice for the nucleus decoding

Red Token Coverage on MSD (Paraphrasis) ↓

Original Text		0.44					
Anita	**Greedy**	0.41					
	Nucleus	0.39					
		Fixed Weights			**Dynamic Weights**		
		Soft	Medium	Strong	Soft	Medium	Strong
	Greedy + UGLD-a	0.38	0.39	0.39	0.39	0.40	0.41
	Nucleus + UGLD-a	0.35	0.31	0.24	0.32	0.25	**0.22**
Qwen	**Greedy**	0.44					
	Nucleus	0.43					
		Fixed Weights			**Dynamic Weights**		
		Soft	Medium	Strong	Soft	Medium	Strong
	Greedy + UGLD-a	0.43	0.43	0.43	0.43	0.43	0.43
	Nucleus + UGLD-a	0.41	0.39	0.34	0.40	0.35	**0.29**

Table 2: In gray the best greedy configuration, in blue is the best nucleus configuration. In **bold** the best configuration overall, per-model.

Green Token Coverage on Wikipedia (Free-Text generation) ↑

Anita	**Greedy**	0.39							
	Nucleus	0.39							
		Uniform			**Top-K**			**Re-Normalization**	
		Soft	Medium	Strong	Soft	Medium	Strong	Soft	Medium	Strong
	Greedy + UGLD-t	0.39	0.39	0.39	0.40	0.41	**0.43**	0.40	0.42	0.42
	Nucleus + UGLD-t	0.39	0.39	0.39	0.40	0.40	0.39	0.40	0.41	**0.43**
Qwen	**Greedy**	0.47							
	Nucleus	0.47							
		Uniform			**Top-K**			**Re-Normalization**	
		Soft	Medium	Strong	Soft	Medium	Strong	Soft	Medium	Strong
	Greedy + UGLD-t	0.47	0.47	0.47	0.48	0.48	**0.50**	0.48	**0.50**	**0.50**
	Nucleus + UGLD-t	0.46	0.47	0.46	0.48	0.47	0.47	0.47	0.49	**0.50**

Table 3: In gray the best greedy configuration, in blue is the best nucleus configuration per model. In **bold** the best configuration overall, per-model.

the baseline.

A last remark concerns the hyperparameters setting. As we can see, across models and conditioning scenarios, the **Re-Normalization and Dynamic Weights configurations tend to yield the strongest lexical control**. This suggests that more adaptive strategies, those that dynamically adjust the intervention at each decoding step, are more effective than static priors in shaping lexical choice. Moreover, the **Strong configuration produces the largest coverage shifts**, reflecting the impact of higher intervention strength.

Human evaluation. Table 5 reports inter-annotator agreement (Fleiss' κ), calculated separately for the first (S1) and second (S2) sentence (or text) in each evaluated pair, and fluency rates (Fluency %) for paraphrasing and free-text generation under baseline nucleus decoding and the uncertainty-gated counterpart (Nucleus+UGLD-t), for both ANITA and Qwen. All evaluations were conducted on outputs generated under the Soft and Strong conditioning strengths. Fluency percentages correspond to the proportion of 50 sentences/texts, per configuration, for which the majority of annotators answered "Yes" (baseline vs. UGLD).

To test whether UGLD affects fluency, we apply the exact McNemar test on paired binary judgments. This test is designed for matched outputs (baseline vs. UGLD for the same input) and evaluates whether the two systems differ in the proportion of positive judgments, focusing on discordant pairs. Across all settings, differences in fluency between UGLD and Nucleus are small, and all of them are non-significant ($p > 0.05$), indicating that UGLD does not systematically degrade grammatical correctness. Even in the Strong conditioning setting, where lexical coverage shifts are largest, fluency remains statistically comparable to baseline decoding. In some configurations, UGLD is numerically higher, though not significantly so.

As we can see, agreement scores range from low to moderate, consistent with prior work on subjective fluency judgments, even among native speakers, as annotators may vary in whether they focus on strictly grammatical aspects or also consider broader orthographic and stylistic factors. Impor-

Red Token Coverage on MSD (Free-Text generation) ↓							
Anita	**Greedy**	0.39					
	Nucleus	0.38					
		Fixed Weights			**Dynamic Weights**		
		Soft	Medium	Strong	Soft	Medium	Strong
	Greedy + UGLD-a	0.39	0.38	0.39	0.38	0.37	0.38
	Nucleus + UGLD-a	0.35	0.28	0.20	0.31	0.19	**0.17**
Qwen	**Greedy**	0.37					
	Nucleus	0.36					
		Fixed Weights			**Dynamic Weights**		
		Soft	Medium	Strong	Soft	Medium	Strong
	Greedy + UGLD-a	0.36	0.36	0.36	0.37	0.37	0.38
	Nucleus + UGLD-a	0.33	0.27	0.17	0.32	0.19	**0.12**

Table 4: In gray the best greedy configuration, in blue is the best nucleus configuration.

			Fleiss κ S1	Fleiss κ S2	UGLD Fluency %	Nucleus Fluency %	p-value
Anita	Paraphrasis	Soft	0.29	0.39	62	72	0.42
		Strong	0.37	0.54	56	54	1.0
	Free-text	Soft	0.2	0.3	82	78	0.80
		Strong	0.36	0.27	44	32	0.28
Qwen	Paraphrasis	Soft	0.43	0.45	68	76	0.57
		Strong	0.28	0.37	84	78	0.64
	Free-text	Soft	0.19	0.15	66	62	0.83
		Strong	0.33	0.20	62	60	1

Table 5: Results of the human evaluation.

tantly, agreement levels are similar across configurations, suggesting that UGLD does not introduce additional instability in perceived fluency. These findings seem to provide empirical support for our objective of developing a decoding-time conditioning approach that enhances lexical control without compromising fluency.

6. Conclusion

In this paper, we introduced Uncertainty-Gated Lexical Decoding (UGLD), a decoding-time framework for fine-grained lexical control in LLMs that explicitly addresses the trade-off between controllability and fluency. Unlike prior decoding-based control approaches, UGLD does not rely on additional training or external discriminators, making it particularly suitable for low-resource settings and languages. Instead, it leverages the entropy of the model's predictive distribution to adaptively scale lexical intervention, activating control when uncertainty is high and limiting interference when predictions are confident. This uncertainty-aware mechanism, combined with explicit lexical priors and logit-level penalties, constitutes the core novelty of our approach.

We evaluated the approach on two open-weight LLMs: ANITA, fine-tuned for Italian, and Qwen 3, a multilingual model supporting Italian generation. The twofold evaluation highlights complementary strengths of UGLD. On the one hand, automatic results show consistent improvements in lexical coverage, both in terms of increased use of simple vocabulary and reduced presence of expert-specific jargon in UGLD-based generations. The analysis further indicates that hyperparameter selection plays a crucial role: configurations that dynamically adapt the intervention at each decoding step and employ stronger intervention strengths produce the largest coverage shifts. On the other hand, human evaluation confirms that these lexical shifts do not negatively affect fluency. This is a particularly relevant outcome, as fluency degradation is a well-known limitation of decoding-time control approaches. Notably, annotators did not perceive a significant decline in fluency even under strong conditioning settings. Overall, these findings suggest that UGLD can serve as a reliable building block for future Controlled Text Simplification frameworks.

7. Acknowledgements

This work has been supported by the project "XAI-CARE" funded by the European Union - Next Generation EU - NRRP M6C2 "Investment 2.1 Enhancement and strengthening of biomedical research in the NHS" (PNRRMAD-2022-2376692_VADALA' – CUP F83C22002470001) and by LLMs4EU "Large

Language Models for the European Union" project, funded by the European Union through the Digital Europe Programme (DIGITAL-2024-AI-B-06-LANGUAGE - GA 101198470) under the grant agreement 101198470.
Partial support was also provided by the project "Understanding and Enhancing Preference Alignment in Large Language Models Through Controlled Text Generation" (IsCc8_ALIGNLLM), funded by CINECA under the ISCRA initiative, for HPC resource availability and support.

8. Lay Summary

Large Language Models (LLMs) are increasingly used to automatically generate text in a wide range of contexts. In many cases, however, it is important to control the type of language they produce. For instance, texts may need to use simpler and more accessible vocabulary, or avoid technical terminology when addressing non-expert readers. Achieving this type of control is challenging, as stronger constraints often reduce the fluency and naturalness of the generated text.

In this work, we propose a method that enables more precise control over the words used by these models while preserving the quality of the generated text. Our approach operates during the text generation process and does not require additional training resources. The key idea is to adjust the level of control depending on how uncertain the model is on what words to generate: stronger guidance is applied when the model is less certain, while more freedom is allowed when it is more confident.

We evaluate the method on Italian using two LLMs: ANITA, which is fine-tuned for Italian, and Qwen 3, a multilingual model capable of generating Italian text. The experiments focus on two scenarios: encouraging the use of simple vocabulary and reducing the presence of domain-specific jargon. In both cases, the method relies on external vocabularies, such as a list of common Italian words and a lexicon of technical terms, to guide the generation.

The results show that the proposed approach effectively increases the use of accessible words and reduces technical language, while maintaining fluency according to human evaluation. These findings suggest that the method can support applications aimed at improving the accessibility and readability of automatically generated texts.

9. Bibliographical References

Souvik Das, Lifeng Jin, Linfeng Song, Haitao Mi, Baolin Peng, and Dong Yu. 2025. Entropy guided extrapolative decoding to improve factuality in large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6589–6600, Abu Dhabi, UAE. Association for Computational Linguistics.

Tullio De Mauro and I Chiari. 2016. Il nuovo vocabolario di base della lingua italiana. *Internazionale*.

Aaron Grattafiori, Abhimanyu Dubey, and et al. 2024. The llama 3 herd of models.

Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1546, Vancouver, Canada. Association for Computational Linguistics.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2021. GeDi: Generative discriminator guided sequence generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4929–4952, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021. DExperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6691–6706, Online. Association for Computational Linguistics.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. 2023. Locally typical sampling. *Transactions of the Association for Computational Linguistics*, 11:102–121.

Dang Nguyen, Jiuhai Chen, and Tianyi Zhou. 2024. Multi-objective linguistic control of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4336–4347, Bangkok, Thailand. Association for Computational Linguistics.

Damian Pascual, Beni Egressy, Clara Meister, Ryan Cotterell, and Roger Wattenhofer. 2021. A plug-and-play method for controlled text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3973–3997, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marco Polignano, Pierpaolo Basile, and Giovanni Semeraro. 2024. Advanced natural-based interaction for the italian language: Llamantino-3-anita.

Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1314–1324, New Orleans, Louisiana. Association for Computational Linguistics.

Seungmin Shin, Dooyoung Kim, and Youngjoong Ko. 2025. ECO decoding: Entropy-based control for controllability and fluency in controllable dialogue generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28297–28309, Suzhou, China. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report.

Kevin Yang and Dan Klein. 2021. FUDGE: Controlled text generation with future discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3511–3535, Online. Association for Computational Linguistics.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2023. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Comput. Surv.*, 56(3).

Tianqi Zhong, Quan Wang, Jingxuan Han, Yongdong Zhang, and Zhendong Mao. 2023. Air-decoding: Attribute distribution reconstruction for decoding-time controllable text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8233–8248, Singapore. Association for Computational Linguistics.

Wangchunshu Zhou, Yuchen Eleanor Jiang, Ethan Wilcox, Ryan Cotterell, and Mrinmaya Sachan. 2023. Controlled text generation with natural language instructions. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

A. Appendix

Table 6 shows an example in paraphrasing decoded with the Nucleus baseline and all the Nucleus+UGLD-t configurations. Words belonging to the New Basic Italian Vocabulary (NBIV) are highlighted in green.

As can be seen, the proportion of green words increases as the conditioning strength increases. For Anita, we can see that Anita Soft rewrites the sentence with the same number of green tokens as the Original, having a coverage lower than the Nucleus counterpart. For all the remaining Anita's settings and all Qwen's settings, higher coverage is achieved than in both the original sentence and the corresponding nucleus settings.

Curiously, while for Qwen the semantic content of the sentence is always correctly preserved during the paraphrasing, Anita either removes or changes the Proper noun of the sentence, from "Dario Arellano" to "Ennio Morricone". However, since this behavior also occurs in the baseline Nucleus setting, it appears to be model-specific rather than induced by UGLD.

Variant	Sentence (NVDB highlighted)	Coverage
Original	Figlio del noto produttore musicale Darío Arellano, fin da piccolo si dedica al teatro e alla recitazione. (lit. *Son of the well-known music producer Darío Arellano, from an early age he devotes himself to theatre and acting.*)	0.294
Nucleus Anita	Figlio del noto compositore musicale Ennio Morricone, da bambino si dedica al teatro e all'attoreplay. (lit. *Son of the well-known musical composer Ennio Morricone, as a child he devotes himself to theatre and to attoreplay*.)	0.312
Nucleus Qwen	Figlio del celebre musicista Darío Arellano, già da bambino si appassiona al teatro e alla recitazione. (lit. *Son of the celebrated musician Darío Arellano, already as a child he becomes passionate about theatre and acting.*)	0.188
Anita Soft	Figlio del noto compositore musicale Ennio Morricone, sin da bambino si dedica al teatro e alla recitazione. (lit. *Son of the well-known musical composer Ennio Morricone, from childhood he devotes himself to theatre and acting.*)	0.294
Anita Medium	Figlio del noto compositore musicale Ennio Morricone, fin da bambino si applica al teatro e all' attore . (lit. *Son of the well-known musical composer Ennio Morricone, from childhood he applies himself to theatre and to the actor.*)	**0.412**
Anita Strong	Figlio del famoso compositore musicale , fin da bambino si appassiona al palcoscenico e all' arte d' attore . (lit. *Son of the famous musical composer, from childhood he becomes passionate about the stage and the art of the actor.*)	**0.353**
Qwen Soft	Figlio del celebre musicista Darío Arellano, da giovanissimo si appassiona al teatro e alla recitazione. (lit. *Son of the celebrated musician Darío Arellano, at a very young age he becomes passionate about theatre and acting.*)	**0.200**
Qwen medium	Figlio del famosissimo musicista Darío Arellano da giovane si affaccia al teatro e alla recitazione. (lit. *Son of the very famous musician Darío Arellano, as a young man he approaches theatre and acting.*)	**0.267**
Qwen Strong	Figlio del celebre musicista Darío Arellano, fin da quando era un bambino si impegnava nel teatro e nello studio di recitazione. (lit. *Son of the celebrated musician Darío Arellano, from when he was a child he was engaged in theatre and in the study of acting.*)	**0.333**

Table 6: Example from Wikipedia in the paraphrasing scenario, generated with standard Nucleus sampling and with each Nucleus+UGLD-t settings. In **bold** all the settings that reach a higher coverage than its corresponding Nucleus baseline or Original, whichever is higher.

A Comparative Study of Multilingual Fine-tuning and Prompting for Automatic Text Readability Classification in Galician

Sandra Rodríguez Rey, Marcos Garcia
Centro Singular de Investigación en Tecnoloxías Intelixentes, Universidade de Santiago de Compostela
Rúa Jenaro de la Fuente Domínguez, 15782, Santiago de Compostela
{sandrarodriguez.rey,marcos.garcia.gonzalez}@usc.gal

Abstract

Despite advancements in automatic readability assessment, low-resource languages such as Galician remain under-explored. This study addresses this gap by presenting a comparative study of readability assessment techniques in Galician, including fine-tuning of encoder models as well as prompting strategies using large generative models. Due to the scarcity of native Galician resources, neural machine translation was employed to generate synthetic Galician data. The analysis begins with BERT-based monolingual models trained on the synthetic data. For multilingual models, the impact of using original versus translated data was compared in order to assess the effects of translation-based augmentation. Finally, several LLMs were evaluated using zero-shot and few-shot prompting methods. The results indicate that generative models are not yet competitive with encoder models tuned for text classification in Galician, and that data generated through machine translation improves the performance of monolingual models but has little effect on multilingual models.

Keywords: Automatic Readability Assessment, Galician, Data Augmentation, Machine Translation, Text Classification.

1. Introduction

Various strategies are currently employed in the development of automatic text classifiers, from linguistically-informed machine learning models to deep learning, hybrid approaches, and strategies involving Large Language Models (LLMs). Recent work indicates that model adjustment using fine-tuning techniques remains one of the most widely used and recommended approaches for multi-class classification tasks (Trokhymovych et al., 2024; Imperial et al., 2025). Concurrently, recent advances in generative artificial intelligence have inspired the exploration of using large language models for classification tasks through instructions and examples (Pungeršek et al., 2025; Kostina et al., 2025).

For similar tasks, and particularly for languages with limited data, some studies suggest that certain data augmentation strategies can improve results (Ziyaden et al., 2024). However, other studies, especially those involving automatically translated data, reveal that the results obtained are slightly inferior to those obtained with original texts (Rehan et al., 2023). Nevertheless, the impact of using translated data may vary depending on the type of model used, for example, monolingual versus multilingual, therefore motivating a focused analysis of this strategy.

In this context, this paper presents a comparative study of strategies to develop automatic text classifiers by readability level in Galician, including original data for this language as well as data augmentation techniques. To this end, this study uses datasets that include texts categorized into four distinct readability levels. These datasets cover multiple textual genres and are specifically designed for developing and evaluating automatic classifiers. The study focuses on three areas: the use of fine-tuned Transformers models, both monolingual and multilingual; the analysis of the impact of using data from other languages, both in their original version and automatically translated into Galician; and the evaluation of large language models provided with instructions and examples.

Thus, the main contributions of this work are as follows: (i) A comparative study of strategies for developing Galician automatic text classifiers by readability level, including fine-tuned and large language models, (ii) an analysis of the impact of data augmentation strategies using corpora from other languages in their original form and in an automatically translated version in Galician in both monolingual and multilingual models, and (iii) the publication of the best models in open repositories, such as Hugging Face[1].

2. Related Work

Early work on readability assessment relied on formula-based metrics grounded in surface features such as sentence length and word complexity (Flesch, 1948; Harry and Laughlin, 1969; Dale and Chall, 1948), followed by mixed feature approaches leveraging tools like Coh-Metrix to capture richer linguistic indicators (Crossley et al., 2007), and more

[1] https://huggingface.co/sandrarrey

Proceedings of the Joint Workshop on Readability and Text Simplification (READIxTSAR) @ LREC 2026, pages 101–120
11 May 2026.

recently by supervised machine learning and deep learning architectures for automatic assessment (Dell'Orletta et al., 2014; Madrazo Azpiazu and Pera, 2019).

Several recent studies on classifying texts according to their readability at multiple levels indicate that fine-tuning pre-trained models remains one of the most widespread and effective approaches (Trokhymovych et al., 2024; Imperial et al., 2025; Ribeiro et al., 2024; Coban et al., 2024; Vásquez-Rodríguez et al., 2022; Hernandez et al., 2022). In practice, this approach constitutes the dominant method for automatically classifying texts by readability level. In parallel, recent advances in generative artificial intelligence have motivated the exploration of using large language models for classification tasks with instructions and examples in the prompt (Pungeršek et al., 2025; Kostina et al., 2025). However, various studies conclude that fine-tuned models offer clearly superior performance to generative models in multi-class classification tasks (Imperial et al., 2025; Pungeršek et al., 2025). Other studies qualify this conclusion, pointing out that the results obtained with generative models can be competitive, although often at the expense of higher computation times (Kostina et al., 2025).

Using corpora annotated by readability levels is essential for training and evaluating these systems. Related work has collected reference corpora and datasets in various languages (Schwarm and Ostendorf, 2005; Wilkens et al., 2022; Ribeiro et al., 2024; Quispesaravia et al., 2016). The iRead4Skills dataset is a recent resource that stands out. It is a multilingual dataset with texts classified by complexity levels in Spanish, Portuguese, and French (Pintard et al., 2024). For languages with limited resources, some studies have been designed to facilitate text evaluation in languages with less available annotated data, such as Basque (Gonzalez-Dios et al., 2018) or Galician (Rodríguez Rey and Garcia, 2025).

In contexts with limited corpus availability, data augmentation strategies are a fundamental resource for improving classifier performance. One of the most widely studied techniques is reusing datasets from other languages, either unchanged or adapted through machine translation. This is done both to obtain readability corpora for low-resource languages (Sibeko, 2024) and to enhance text classification tasks in general (Ziyaden et al., 2024). While some studies show that models trained with machine translated augmented data can significantly outperform those trained with only original data (Ziyaden et al., 2024), other studies conclude that translated texts generate slightly lower performance in models compared to original texts (Rehan et al., 2023). This apparent contradiction in results shows that the effectiveness of this strategy depends on the type of model used, the quality of the translation, and the specific language. Therefore, it is necessary to investigate how combining data augmentation strategies with monolingual or multilingual approaches, such as mBERT or XLM-RoBERTa, affects the performance of classifiers in languages with few resources (Pakray et al., 2025).

Automatic readability assessment plays a vital role in language education and digital accessibility (Vajjala, 2022). However, despite its recognized significance for learning and inclusion, notable gaps remain in the research on text readability in languages with limited resources. While the broader field has seen recent progress, strategies for developing high-quality automatic classifiers for these languages have hardly been analyzed. For Galician, building robust models requires a sufficient amount of annotated data. However, only a small corpus is available for this task (Rodríguez Rey and Garcia, 2025). This makes reusing multilingual datasets and exploring data augmentation techniques necessary. This is where this work comes in.

This study aims to advance the development of automatic text classifiers for Galician by comparing different models and analyzing the impact of using translated versus original texts in monolingual and multilingual scenarios. Beyond a technical perspective, this work aims to contribute to the design of specialized pedagogical tools and to foster the creation and adaptation of educational materials for Galician language teaching. One practical application of these classifiers, for example, is that they could help teachers and independent learners easily find and select reading materials that are just right for their level. This is particularly important for languages that have recently been incorporated into formal education, since even native adult speakers may have difficulty achieving proficiency and developing consistent reading habits in Galician.

3. Research Questions (RQs) and Hypotheses (Hs)

Building on prior studies which show that fine-tuning encoder models is effective for multi-class text classification (Trokhymovych et al., 2024; Pungeršek et al., 2025; Imperial et al., 2025), and that multilingual models and data augmentation improve performance (Ziyaden et al., 2024), we formulate the following research questions and corresponding hypotheses:

- RQ1: Does translation-based data augmentation improve the performance of fine-tuned models?

H1: Monolingual models would benefit most from the introduction of data in the target language. Multilingual models could perform better with a larger volume of texts translated into the same language in which they will be evaluated. However, if the languages of the original resources are included in the multilingual model, translation may not be necessary.

- RQ2: Considering the recent advancements in generative models, is it worthwhile to fine-tune a model for multi-class text classification, or can generative models achieve competitive or even superior performance to fine-tuned models?
  H2: As indicated by related work (Imperial et al., 2025), model fine-tuning remains the most effective technique for multi-class classification tasks. However, recent advances in generative AI, especially instructed models, could compete with this technique (Pungeršek et al., 2025; Kostina et al., 2025), as these models can work with instructions and few examples, saving work.

4. Resources: datasets and models

The following datasets and language models were used to conduct experiments comparing the performance of different tools for classifying texts according to their readability in Galician.

4.1. Datasets

The following corpora were utilized. A proprietary script was used to convert them into CSV datasets compatible with the libraries used for model development.

Corlega corpus[2] : This corpus includes 480 Galician texts (145,854 tokens) from various subgenres and themes, such as social media, fiction and nonfiction literature, professional websites, ads, political discourse, and legal texts, extracted from diverse online sources. The texts are aimed at adult Galician speakers looking to improve their skills and learners of Galician. One expert annotator classified the documents into four levels (first to fourth). These levels were defined by a set of linguistic descriptors describing the lexical-conceptual, verbal, syntactic, cohesion, and textual characteristics of each level. The levels are based on those defined in the iRead4Skills multilingual dataset (see below) and have been adapted primarily according to the specifications of the Celga (the standard system for certifying proficiency in Galician) and CEFR (Common European Framework of Reference for Languages) levels (Rodríguez Rey and Garcia, 2025).

iRead4Skills Dataset 1: This dataset includes three corpora of Spanish, Portuguese, and French texts classified by complexity (Pintard et al., 2024). These corpora consist of 2,000 to 3,000 texts per language from various subgenres and themes intended for native adult speakers. The texts are classified into four levels (very easy, easy, clear, and more complex), three of which are defined by experts (levels one to three), and a fourth reference level that includes texts that are more complex than the previous ones (Monteiro et al., 2024).

Galician translation of the iRead4Skills Dataset 1: The dataset previously mentioned was automatically translated into Galician using SalamandraTA[3], and this translated version was also used in some experiments.

4.1.1. Datasets splitting

The Corlega corpus was used for the evaluation, and the iRead4Skills Dataset 1 and its Galician translation were used for training. The Corlega corpus was divided into two parts: validation (25%) and evaluation (75%). Using a proprietary script, the texts were chosen randomly to ensure a balance in the number of texts per level in each part. Given the imbalance in the number of texts per level in the training sets, the final amount was adjusted using an oversampling strategy. Texts were randomly duplicated from levels with fewer documents to equalize the model's recognition capacity across all levels. An automated consistency check was applied to the translated dataset, and documents were excluded when the length ratio of the original and translated texts differed by more than 30%. In total, 119 translated texts were deleted, accounting for 1.56% of the texts. The final size of the datasets is shown in Table 1.

4.2. Models

The following models, both encoders and LLMs, were used:

4.2.1. Encoders

The following monolingual and multilingual models, extracted from the Transformers library in Hugging

[2]The Corlega corpus is available under a CC BY-NC-ND 4.0 license for research and reproducibility purposes at Zenodo (Rodríguez Rey and Garcia, 2025). More information about corpus creation is available at Rodríguez Rey and Garcia (2025).

[3]https://huggingface.co/BSC-LT/salamandraTA-7b-instruct

Language	split	O.T.	T.T.
GL	Test	360	-
GL	Valid	120	-
PT+SP+FR	Train	7677	9728
GL(MT)*	Train	7558	9556

Table 1: Total number of texts in the original datasets *(O.T.)* and total amount of texts with oversampling *(T.T.)*. *GL(MT) refers to the PT+SP+FR dataset translated into Galician.

Face (Wolf et al., 2020), were used as a starting point to design a baseline model and perform the fine-tuning experiments:

- BERT-gl (Garcia, 2021), using its two variants: *small* and *base*.
- Bertinho (Vilares et al., 2021), also in its *small* and *base* variants.
- mBERT (Devlin et al., 2019), a multilingual BERT-base model.
- XLM-RoBERTa (Conneau et al., 2019) in its *base* and *large* variants.

4.2.2. Generative models

To explore the potential of LLMs for the proposed classification task, the following models were used:

- Llama-3.1-8B-Instruct[4] (Grattafiori et al., 2024). It includes closely related Romance languages, such as Spanish, Portuguese, French, and Italian, but not Galician.
- Llama3.1-Carvalho-PT-GL 8B[5], a continuation of the training of Llama 3.1 8B (Grattafiori et al., 2024) on a large monolingual Galician corpus (de Dios-Flores et al., 2024).
- Gemma-3-4b-it[6] (Gemma Team, 2025), a multilingual model that includes 140 languages and has been trained to perform various text and image generation tasks. Although the list of supported languages is not explicitly stated, it can be assumed that Galician is included because the model was evaluated using the Flores-101 benchmark (Goyal et al., 2022), which includes Galician.
- Qwen2.5-7B-Instruct[7] (Yang et al., 2024; Team, 2024), the best-performing model for Galician in the IberBench Leaderboard (Ángel González et al., 2026)[8].
- GPT-5[9], as an example of state-of-the-art proprietary model, used via ChatGPT[10].

5. Experiments and results

This section introduces a set of baseline experiments followed by the results of the fine-tuned models as well as of the LLMs. The experiments were performed on a standard server equipped with two NVIDIA Hopper H100 GPUs, each with 80 GB of memory, and 384 GB of RAM.

5.1. Baseline

As a baseline for comparison with all other models, a simple hybrid approach which combines surface features and information from Transformer models to train Random Forest classifiers was implemented. As textual features, the mean, median, mininum and maximum sentence and word length of each document were included, totaling 8 features. From Transformers, a document vector, and surprisal were included. The former is constructed by averaging all sentence vectors in the document, each sentence being represented as the mean of all its sub-tokens (Imperial, 2021). Regarding surprisal, the method proposed by Kauf and Ivanova (2023) for bidirectional models was followed, using the implementation provided by `minicons`[11] (Misra, 2022).

The classifiers were trained with the Random Forest implementation of `scikit-learn`[12], using the datasets presented in Section 4.1 and the encoder models mentioned in Section 4.2.1. During training, different estimators sizes (50 to 600) were evaluated and the best configuration based on the performance on the validation set was selected.

The results of the experiments with the baseline models are shown in Table 2. For these baseline models, accuracy for performance comparison is reported.

5.2. Fine-tuning encoder models

To answer RQ1, three types of experiments were conducted to fine-tune the Transformer models

[4] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
[5] https://huggingface.co/Nos-PT/Llama-Carvalho-PT-GL
[6] https://huggingface.co/google/gemma-3-4b-it
[7] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct
[8] https://huggingface.co/spaces/iberbench/leaderboard
[9] https://openai.com/es-ES/index/introducing-gpt-5/
[10] https://chatgpt.com/
[11] https://github.com/kanishkamisra/minicons
[12] https://scikit-learn.org

Model	Dataset	Accuracy
BERT-small	GL(MT)	**0.51**
BERT-base	GL(MT)	0.49
Bertinho-base	GL(MT)	0.47
Bertinho-small	GL(MT)	0.44
XLM-base	PT+SP+FR	*0.50*
mBERT-base	PT+SP+FR	0.48
XLM-large	PT+SP+FR	0.47
XLM-base	GL(MT)	*0.48*
mBERT-base	GL(MT)	*0.48*
XLM-large	GL(MT)	0.46

Table 2: Best results from experiments with baseline models trained with authentic multilingual data (PT+SP+FR) and translated data (GL(MT)), evaluated on the test set. The best results are shown in bold. In italics, the best results in each scenario.

mentioned in Section 4.2.1. The models and training datasets were divided into monolingual and multilingual sets for the different experiments.

Monolingual: These experiments were conducted using the monolingual models in Galician, BERT-gl and Bertinho, fine-tuned using the Galician-translated dataset.

Multilingual with authentic data: In this case, the experiments were carried out using the multilingual encoders (mBERT and XLM-RoBERTa). Adjustments were made using the original dataset, which contained classified texts in Portuguese, Spanish, and French.

Multilingual with translated data: Here the same multilingual models adjusted using the Galician-translated dataset were used.

A standard fine-tuning methodology was employed, consisting in adjusting the model parameters for text classification based on a training corpus. Several hyperparameters related to batch size (8, 16, and 32) and five learning rate values (from 1e-5 to 5e-5) were explored. The models were adjusted with different combinations of these hyperparameters for five epochs and then evaluated using the validation set. The best configurations were selected for each model based on accuracy, which were then evaluated using the test set[13].

Table 3 shows the results of the fine-tuning experiments. The best results for each model were chosen based on accuracy.

[13]Since the results on the validation set were not conclusive for some models (for example, there was a tie for several configurations of the same model), more than one configuration for some models were also evaluated on the test set.

5.3. Zero-shot and few-shot LLM prompting

Various experiments were carried out to compare the performance of generative models' potential for this task, employing a standardized methodology and focusing on answering RQ2. In the following, we first present the generic instruction included in the prompts, and then introduce the zero- and few-shot prompting methods.

Instructions: The prompts were developed following OpenAI's best practices for prompt engineering[14] and were written in Spanish[15]. Each prompt included role, objective, classification level characteristics, step-by-step task instructions, context, expected model inputs, output format and criteria, restrictions, and common errors to avoid. All models were asked to classify input texts into the four described levels. Each prompt specified:

- The description of the four levels, including their lexical-conceptual, syntactic, verbal, and cohesion features, as well as the most frequent text genres associated with each level[16].
- The requirement to analyze the provided examples (when available).
- The task of classifying the texts from a CSV file by assigning a numerical level from 1 to 4.
- The input data: level descriptions, texts to classify, and, in some cases, example texts.
- The expected output: a CSV table containing the predicted level for each text.
- Restrictions and errors to avoid: inventing characteristics and levels, randomly assigning a level, and responding with information that is not strictly a number from one to four.

Zero-shot strategy: In the experiments, only a prompt containing the specified instructions was used, and a set of test texts was provided for level prediction.

[14]https://tinyurl.com/openai-best-practices-prompt

[15]Since LLMs tend to perform better in high-resource languages (Xuan et al., 2025), we conducted these experiments in Spanish, which is significantly more representative than Galician and co-exists with it in the same territory.

[16]This information was extracted directly from the appendix published in Rodríguez Rey and Garcia (2025).

Model	Dataset	Acc	NAcc	Prec	Rec	F1
Bertinho-base	GL(MT)	**0.55**	**0.95**	*0.56*	**0.55**	**0.54**
BERT-base	GL(MT)	0.54	**0.95**	*0.56*	0.49	0.50
Bertinho-small	GL(MT)	0.52	0.94	0.52	0.51	0.52
BERT-small	GL(MT)	0.51	0.94	0.51	0.51	0.51
XLM-base	GL(MT)	*0.54*	0.92	*0.54*	*0.53*	*0.53*
XLM-large	GL(MT)	0.52	0.92	0.51	*0.53*	0.51
mBERT	GL(MT)	0.51	*0.94*	0.52	0.47	0.48
XLM-base	PT+SP+FR	0.53	*0.93*	0.56	*0.50*	*0.52*
XLM-large	PT+SP+FR	*0.54*	0.91	**0.58**	*0.50*	0.50
mBERT	PT+SP+FR	0.50	0.91	0.52	0.46	0.47

Table 3: Best results from fine-tuning monolingual models with translated data (GL(MT)), multilingual models with authentic data (PT+SP+FR), and multilingual models with translated data (GL(MT)). *Acc* refers to accuracy, *NAcc* to neighbor or adjacent accuracy, *Prec* to precision, *Rec* to recall, and *F1* to the F-score obtained in the test set. The best results are shown in bold. In italics, the best results in each scenario.

Few-shot strategy: In this scenario, 13 examples of classified texts were added to the general instructions within the same prompt. The texts were distributed by level as follows: five examples of level 1, four examples of level 2, three examples of level 3, and one example of level 4. The limited number of examples is due to the maximum size allowed for the prompt. The total size of the examples at each level (one to three) is similar. Only one level 4 example was included because this level is not defined by experts; it is simply a reference for texts that are more complex than those of the defined levels.

Table 4 shows the results of the experiments with generative models.

6. Discussion of results and comparative study

Some ideas and trends can be drawn from the results obtained in the different experiments. These results allow us to answer the research questions and to validate or reject the established hypotheses. All calculated metrics were considered in the evaluation, with accuracy serving as the primary reference for selecting the best-performing models.

The baseline models, a hybrid model combining surface features and information extracted from encoder models, achieved similar scores with both original and translated data, as well as with both monolingual and multilingual models. The best results were 0.51 accuracy with a monolingual model and translated data and 0.50 accuracy with a multilingual model and original data. The results suggest that multilingual models perform slightly better with original data, but better results were obtained in only half of the models and the difference was minimal. Additionaly, there are no clear trends regarding the influence of model size on results.

Experiments with fine-tuned models reveal similar trends to those observed with baseline models. There is little difference in the results of monolingual and multilingual models or between models fine-tuned with original or translated data. The best results across all metrics were achieved with a Bertinho base model fine-tuned with translated data. This was followed by a BERT-base model, that had nearly identical precision results but lower recall. This implies that the F1 score is also lower than that of the Bertinho-base model. For multilingual models, the results for mBERT and XLM-base with original and translated data are very similar, though the translated models have a slight advantage. However, XLM-large with original data has better precision (the highest: 0.58), although lower recall than XLM-base with translated data, which is generally the best multilingual model.

Experiments with generative models developed in this study did not reach the minimum values established by baseline models. This seems to indicate that these models, with a standard methodology, are not currently the best option for multi-class text classification tasks. However, more advanced strategies (Liu et al., 2024; Yousefiramandi and Cooney, 2025) can be explored in the future. As expected, the results show that few-shot technique improves the performance of almost all models, except for Llama3.1 Instruct in terms of accuracy and neighbor accuracy, and Gemma-3 in terms of precision. The best model in terms of accuracy with the zero-shot technique is Llama3.1 Instruct, though Gemma-3 is the best model in general terms. Gemma-3 achieves an accuracy and adjacent accuracy slightly below that of Llama, but significantly outperforms it in the rest of the metrics. As anticipated from prior information regarding its high performance in classification tasks, the best model with the few-shot technique is Qwen.

To gain a deeper understanding of how the model

	Zero-shot					Few-shot				
Model	**Acc**	**NAcc**	**Prec**	**Rec**	**F1**	**Acc**	**NAcc**	**Prec**	**Rec**	**F1**
Llama3.1 Instruct	*0.37*	*0.86*	0.29	0.31	0.28	0.35	0.72	0.32	0.44	0.27
Carvalho-PT-GL	0.31	0.82	0.24	0.27	0.23	0.37	0.83	**0.55**	0.32	0.28
Gemma-3	0.36	0.82	*0.42*	*0.37*	*0.34*	0.40	0.84	0.38	0.41	0.35
Qwen2.5 Instruct	0.30	0.74	0.37	0.32	0.27	**0.43**	**0.88**	0.40	**0.45**	**0.41**
GPT-5	0.16	0.58	0.36	0.26	0.13	0.30	0.78	0.28	0.26	0.26

Table 4: Best results from experiments with generative models evaluated on the test set. *Acc* refers to accuracy, *NAcc* to neighbor or adjacent accuracy, *Prec* to precision, *Rec* to recall, and *F1* to the F-score obtained in the test set. The best results are shown in bold. In italics, the best results in each scenario.

behaves at different levels, a confusion matrix for the best-performing model in each fine-tuning experimental scenario was created. For this analysis, the Bert-base model adjusted with translated data, the XLM-base model adjusted with translated data, and the XLM-large model with data in its original language were selected. Information from the models' predictions on the test was used to create the confusion matrices in Figure 1. The conclusion that Bertinho-base is the best-performing model is reinforced by these visualizations, which complement the quantitative metrics. It classified very few texts with a difference of more than one level and showed better prediction of level 4, which had fewer original texts in its adjustment. Of the two multilingual models selected, XLM-base with translated data made the best predictions. XLM-large, when trained on the original dataset, demonstrates a pronounced error concentration in levels 3 and 4.

Table 5 shows in greater detail the differences in performance by level between the two best models, Bertinho-base and XLM-base, which were both fine-tuned with translated data. Bertinho-base outperforms XLM-base on all metrics and levels, with the exception of level 2 predictions, in which XLM-base surpasses Bertinho-base.

L	Bertinho-base			XLM-base		
	Prec	**Rec**	**F1**	**Prec**	**Rec**	**F1**
1	0.70	0.67	**0.68**	0.65	0.62	0.63
2	0.45	0.53	0.49	0.46	0.59	**0.52**
3	0.56	0.46	**0.50**	0.55	0.44	0.49
4	0.48	0.54	**0.51**	0.51	0.46	0.49

Table 5: Results by level for the two best models. L represents the level, *Prec* the precision, *Rec* the recall, and *F1* the F-measure. The best results are shown in bold.

Overall, our experimental results are consistent with the trends observed in previous studies, particularly the marginal advantage of monolingual models. Direct comparison, however, is not possible due to the significantly smaller test set used in the prior work (Rodríguez Rey and Garcia, 2025). After discussing the results, they are then analyzed in conjunction with similar studies to provide a more comprehensive answer to the research questions.

- A1: For monolingual models and limited resources, increasing data by incorporating automatically translated data seems to be beneficial. However, for multilingual models, this strategy is only beneficial in some cases, and the margin of improvement is small. Furthermore, studies for similar tasks obtain slightly lower results with translated texts than with original texts (Rehan et al., 2023). Therefore, considering the time, effort, and computing costs involved, this strategy may not be worthwhile for multilingual models.

- A2: Fine-tuning models to classify texts into multiple classes remains worthwhile because the performance of generative models for this task is far from that of fine-tuned models, as shown by the results of this study and other similar works (Imperial et al., 2025). The same conclusion has been reached in other text classification tasks, such as classifying texts by genre or textual theme (Pungeršek et al., 2025). Furthermore, using language models with fine-tuning techniques achieves state-of-the-art results in numerous natural language understanding tasks (Devlin et al., 2019; Ngo and Parmentier, 2023). Many studies use this technique to classify texts according to their complexity since it is currently the most widespread technique in the field (Trokhymovych et al., 2024; Ribeiro et al., 2024; Coban et al., 2024; Vásquez-Rodríguez et al., 2022; Hernandez et al., 2022).

  However, other studies draw slightly different conclusions. They argue that the results obtained by fine-tuned models are competitive with those of generative models and with much shorter computation times (Kostina et al., 2025). It should also be noted that this study employed a standard methodology and that other approaches exist—such as using the

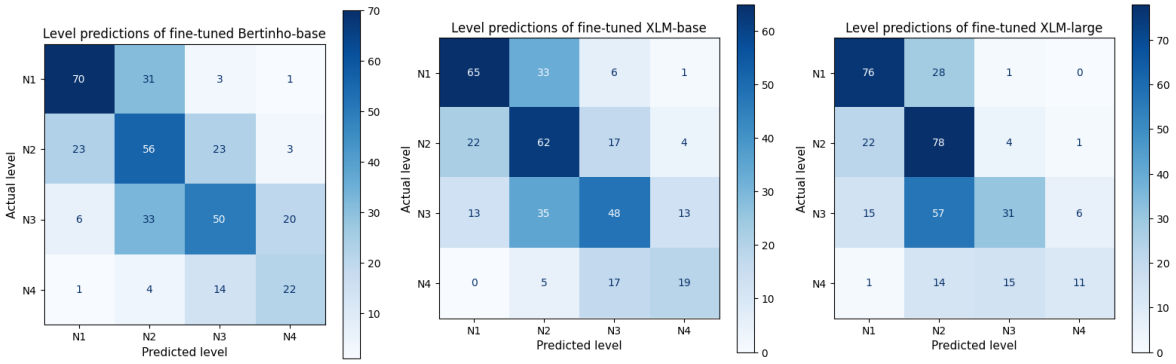


Figure 1: Distribution of predictions from the models with the best results on the test set.

same prompt translated into English, or even more advanced methods, such as using model-generated rubrics (Liu et al., 2024) or fine-tuning LLMs with LoRA (Yousefiramandi and Cooney, 2025)—which could yield better results.

Regarding the use of instructed or uninstructed models, the results obtained do not clearly support Hypothesis 2 (see Section 3) that instructed models perform better on this task. As for zero-shot and few-shot techniques, the results show that the few-shot technique, which includes several examples of the task, achieves better results.

7. Conclusions and further work

This paper have presented a comparative study of strategies for developing automatic text classifiers by readability level in Galician, as well as data augmentation approaches. The primary methods for developing classifiers involve using fine-tuned models (monolingual and multilingual) and large generative language models that are provided with instructions and examples of the task. Data augmentation techniques include using suitable task-specific datasets in other languages and automatically translating them into Galician.

To conduct the comparative study, datasets and models were selected and a series of experiments were carried out to answer the research questions. A baseline model combining linguistic features, vectors, and surprisal was established to evaluate the performance of the different models. Regardless of the type of model or data used, these models obtained similar results, with the best model achieving an accuracy of 0.51 using the translated dataset and a monolingual model.

Fine-tuned models consistently produce equal or superior results in both monolingual and multilingual experiments. Considering all the analyzed metrics, the best result is obtained with the Bertinho base model and translated data. Other models achieve slightly lower results, including both monolingual and multilingual models with both original and translated data. In contrast, generative models using a standard methodology, have shown lower performance compared to baseline models. In general, generative models provided with examples (few-shot) perform better than those provided only with instructions.

This study evaluated the results of different experiments and other similar works and concluded that, in the case of Galician, the strategy of data augmentation through machine translation of corpora designed for the task in other languages is beneficial when using monolingual models. For multilingual models, however, the margin for improvement may be minimal or nonexistent. Regarding generative models' performance in classifying Galician texts by readability level, it is concluded that, with a standard methodology, for now, they are not competitive with fine-tuned models.

Future work includes expanding the Galician readability corpus by increasing the number of texts per readability level and extending the range of levels covered. Further research may explore more sophisticated prompting strategies with LLMs and fine-tuning approaches beyond encoder-based architectures. These efforts aim to assess how data augmentation, cross-lingual transfer, and model design impact the automatic readability assessment of texts in Galician.

8. Limitations

Regarding generative models, the experiments were limited to the standard methodology of using two nearly identical prompts for the zero-shot and few-shot strategies. A more diverse approach would have involved testing prompts in different languages, such as Galician, Spanish, and English. Additionally, more advanced strategies could have been explored, such as model-generated rubrics, to

better understand AI-specific reasoning compared to the human-defined criteria used in this study. Other strategies that could have been explored include fine-tuning via LoRA. Regarding the data, the evaluation corpus was relatively small (480 texts), and the translated corpus lacked manual supervision. Lastly, while we tested several architectures, using a wider variety of Transformers and LLMs could have strengthened the generalizability of the findings.

9. Acknowledgements

This work was supported by the European Commission (Project: iRead4Skills, Grant number: 1010094837, Topic: HORIZON-CL2-2022-TRANSFORMATIONS-01-07, DOI: 10.3030/101094837), by MCIU/AEI/10.13039/501100011033 (grants with references PID2021-128811OA-I00, CNS2024-154902, PID2024-161928OB-I00, and AIA2025-163322-C62), and by the Galician Government (ERDF 2024-2027: Call ED431G 2023/04, and ED431B 2025/16).

10. Lay Summary

Automatic tools that measure how difficult a text is to read are essential for language learning and making information accessible to everyone. While these tools are common for major languages, they are still rare for languages with fewer digital resources, like Galician. This study addresses this gap by testing different artificial intelligence methods to see which works best for classifying Galician texts by their complexity. Since the available native Galician dataset for this specific task is very limited in size, we used machine translation to create a larger set of training materials for the AI. We then compared two main strategies: training specialized models specifically for this task and using large, general-purpose AI models by giving them clear instructions and examples. Our results show that the specialized models are currently much more effective than the general AI models at accurately judging text difficulty in Galician. Furthermore, we found that using translated data significantly helps models that focus solely on Galician, whereas it has a limited impact on models that are already trained in multiple languages.

11. Bibliographical References

Onder Coban, Mete Yağanoğlu, and Ferhat Bozkurt. 2024. Domain effect investigation for bert models fine-tuned on different text categorization tasks. *Arabian Journal for Science and Engineering*, 49(3):3685–3702.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Scott Andrew Crossley, David F. Dufty, Philip M. McCarthy, and Danielle S. McNamara. 2007. Toward a new readability: A mixed model approach. In *Proceedings of the annual meeting of the cognitive science society*, volume 29.

Edgar Dale and Jeanne Sternlicht Chall. 1948. *A formula for predicting readablility*. Bureau of Educational Research, Ohio State University.

Iria de Dios-Flores, Silvia Paniagua Suárez, Cristina Carbajal Pérez, Daniel Bardanca Outeiriño, Marcos Garcia, and Pablo Gamallo. 2024. CorpusNÓS: A massive Galician corpus for training large language models. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 593–599, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2014. Assessing document and sentence readability in less resourced languages and across textual genres. *ITL - International Journal of Applied Linguistics*, 165:163–193.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Rudolf Flesch. 1948. A readability formula in practice. *Elementary English*, 25(6):344–351.

Marcos Garcia. 2021. Exploring the representation of word meanings in context: A case study on homonymy and synonymy. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3625–3640, Online. Association for Computational Linguistics.

Gemma Team. 2025. Gemma 3.

José Ángel González, Ian Borrego Obrador, Álvaro Romo Herrero, Areg Mikael Sarvazyan, Mara Chinea-Ríos, Angelo Basile, and Marc Franco-Salvador. 2026. Iberbench: Llm evaluation on iberian languages. *Computer Speech & Language*, 96:101899.

Itziar Gonzalez-Dios, María Jesús Aranzabe, and Arantza Díaz de Ilarraza. 2018. The corpus of basque simplified texts (cbst). *Language Resources and Evaluation*, 52(1):217–247.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Aaron Grattafiori, et al., 2024. The Llama 3 Herd of Models.

G Harry and Mc Laughlin. 1969. Smog grading - a new readability formula. *The Journal of Reading*.

Nicolas Hernandez, Nabil Oulbaz, and Tristan Faine. 2022. Open corpora and toolkit for assessing text readability in French. In *Proceedings of the 2nd Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI) within the 13th Language Resources and Evaluation Conference*, pages 54–61, Marseille, France. European Language Resources Association.

Carina Kauf and Anna A. . 2023. A better way to do masked language model scoring. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 925–935, Toronto, Canada. Association for Computational Linguistics.

Arina Kostina, Marios D. Dikaiakos, Dimosthenis Stefanidis, and George Pallis. 2025. Large language models for text classification: Case study and comprehensive review.

Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. 2024. Calibrating LLM-Based Evaluator. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2638–2656.

Ion Madrazo Azpiazu and Maria Soledad Pera. 2019. Multiattentive recurrent neural network architecture for multilingual readability assessment. *Transactions of the Association for Computational Linguistics*, 7:421–436.

Joseph Marvin Imperial. 2021. BERT embeddings for automatic readability assessment. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 611–618, Held Online. INCOMA Ltd.

Joseph Marvin Imperial, Abdullah Barayan, Regina Stodden, Rodrigo Wilkens, Ricardo Muñoz Sánchez, Lingyun Gao, Melissa Torgbi, Dawn Knight, Gail Forey, Reka R. Jablonkai, Ekaterina Kochmar, Robert Joshua Reynolds, Eugénio Ribeiro, Horacio Saggion, Elena Volodina, Sowmya Vajjala, Thomas François, Fernando Alva-Manchego, and Harish Tayyar Madabushi. 2025. UniversalCEFR: Enabling open multilingual research on language proficiency assessment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9703–9755, Suzhou, China. Association for Computational Linguistics.

Kanishka Misra. 2022. minicons: Enabling flexible behavioral and representational analyses of transformer language models. *arXiv preprint arXiv:2203.13112*.

Ricardo Monteiro, Raquel Amaro, Susana Correia, Alice Pintard, Roser Gauchola, Michell Moutinho, and Xavier Blanco Escoda. 2024. iRead4Skills - Complexity Levels.

Duy Van Ngo and Yannick Parmentier. 2023. Towards sentence-level text readability assessment for French. In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 78–84, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Partha Pakray, Alexander Gelbukh, and Sivaji Bandyopadhyay. 2025. Natural language processing applications for low-resource languages. *Natural Language Processing*, 31(2):183–197.

Taja Kuzman Pungeršek, Peter Rupnik, Ivan Porupski, Vuk Dinić, and Nikola Ljubešić. 2025. State of the art in text classification for south slavic languages: Fine-tuning or prompting?

Andre Quispesaravia, Walter Perez, Marco Sobrevilla Cabezudo, and Fernando Alva-Manchego. 2016. Coh-Metrix-Esp: A complexity analysis tool for documents written in Spanish. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4694–4698, Portorož, Slovenia. European Language Resources Association (ELRA).

Muhammad Rehan, Muhammad Shahid Iqbal Malik, and Mona Mamdouh Jamjoom. 2023. Fine-tuning transformer models using transfer learning for multilingual threatening text identification. *IEEE Access*, 11:106503–106515.

Eugénio Ribeiro, Nuno Mamede, and Jorge Baptista. 2024. Automatic text readability assessment in European Portuguese. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 97–107, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Sandra Rodríguez Rey and Marcos Garcia. 2025. Clasificación automática de textos por niveis de lecturabilidade: recursos e modelos para o galego. *Linguamática*, 17(2):33–56.

Sarah Schwarm and Mari Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 523–530, Ann Arbor, Michigan. Association for Computational Linguistics.

Johannes Sibeko. 2024. Harnessing google translations to develop a readability corpus for sesotho: An exploratory study. *Journal of the Digital Humanities Association of Southern Africa (DHASA)*, 5(1).

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Mykola Trokhymovych, Indira Sen, and Martin Gerlach. 2024. An open multilingual system for scoring readability of Wikipedia. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6296–6311, Bangkok, Thailand. Association for Computational Linguistics.

Sowmya Vajjala. 2022. Trends, Limitations and Open Challenges in Automatic Readability Assessment Research. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 5366–5377.

Laura Vásquez-Rodríguez, Pedro-Manuel Cuenca-Jiménez, Sergio Morales-Esquivel, and Fernando Alva-Manchego. 2022. A benchmark for neural readability assessment of texts in Spanish. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 188–198, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

David Vilares, Marcos Garcia, and Carlos Gómez-Rodríguez. 2021. Bertinho: Galician bert representations. *Procesamiento del Lenguaje Natural*, 66(0):13–26.

Rodrigo Wilkens, David Alfter, Xiaoou Wang, Alice Pintard, Anaïs Tack, Kevin P. Yancey, and Thomas François. 2022. FABRA: French aggregator-based readability assessment toolkit. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1217–1233, Marseille, France. European Language Resources Association.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, Jinghui Lu, Yuang Jiang, Huitao Li, Xin Li, Kunyu Yu, Ruihai Dong, Shangding Gu, Yuekang Li, Xiaofei Xie, Felix Juefei-Xu, Foutse Khomh, Osamu Yoshie, Qingyu Chen, Douglas Teodoro, Nan Liu, Randy Goebel, Lei Ma, Edison Marrese-Taylor, Shijian Lu, Yusuke Iwasawa, Yutaka Matsuo, and Irene Li. 2025. MMLU-ProX: A Multilingual Benchmark for Advanced Large Language Model Evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 1513–1532.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan.

2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Amirhossein Yousefiramandi and Ciaran Cooney. 2025. Fine-Tuning Causal LLMs for Text Classification: Embedding-Based vs. Instruction-Based Approaches. *arXiv preprint arXiv:2512.12677*.

A Ziyaden, A Yelenov, F Hajiyev, S Rustamov, and A Pak. 2024. Text data augmentation and pretrained language model for enhancing text classification of low-resource languages. *PeerJ Computer Science*, 10:e1974.

12. Language Resource References

Alice Pintard, Thomas François, Justine Nagant de Deuxchaisnes, Sílvia Barbosa, Maria Leonor Reis, Michell Moutinho, Ricardo Monteiro, Raquel Amaro, Susana Correia, Sandra Rodríguez Rey, Marcos Garcia González, Keran Mu, and Xavier Blanco Escoda. 2024. iRead4Skills Dataset 1: corpora by complexity level for FR, PT and SP.

Sandra Rodríguez Rey and Marcos Garcia. 2025. Corlega: Corpus de Lecturabilidade en Galego.

13. Appendix A. Prompt used in LLM experiments

[Rol] Actúa como un lingüista experto en gallego, especializado en clasificación de textos por complejidad. Eres experto en clasificar textos en los niveles de complejidad definidos a continuación. Respondes a los textos de entrada que aparecen en el siguiente mensaje indicando claramente a qué nivel pertenecen.

[Objetivo] Tu objetivo es clasificar textos de entrada según su complejidad en cuatro niveles: tres niveles definidos y un cuarto nivel no definido que incluye textos más complejos que los definidos. Las clasificaciones deben ser claras y solamente en uno de los niveles. Los niveles se indicarán solamente con un número del 1 al 4, correspondiendo el 1 al nivel 1, el 2 al nivel 2, el 3 al nivel 3 y el 4 al nivel 4. Contarás con una serie de ejemplos de cada nivel para que entiendas mejor la tarea. Estas son las características de los diferentes niveles:

- Nivel 1: Incluye textos cotidianos, como mensajes cortos de correo o relacionados con las actividades diarias, letreros, listados, manuales, propaganda, formularios, materiales relacionados con sus intereses, documentos auténticos breves (billetes, entradas, cartas de restaurante, facturas, etiquetas, planos, embalajes, horarios, mapas...) de uso muy frecuente, instrucciones sencillas, relatos fáciles, letras de canciones y poemas sencillos. Incluye los adverbios más comunes de cantidad (moito, pouco), tiempo (hoxe, agora, cedo), lugar (enriba, lonxe, preto), modo (ben, mellor, amodo), exclusión (só, soamente), afirmación y negación. Características Léxico-conceptuales: Nombres comunes y propios. Sufijo- iño/a. Valor de los afijos más frecuentes. Superlativo en- ísimo/a. Repertorio de vocabulario limitado a situaciones cotidianas y de especial interés. Características verbales: Verbos copulativos. Verbos regulares y semirregulares de uso frecuente (durmir, servir, espir). Irregulares más frecuentes (ir, facer, estar). Reflexivos y pronominales más comunes (chamarse, esquecerse, queixarse). Tiempos de indicativo en presente, pasado y futuro. Futuro hipotético. Perífrases más habituales: ir + infinitivo, estar/andar + gerundio, volver + infinitivo, comezar a + infinitivo, ter que + infinitivo. Características de estructura sintáctica: Coordinadas simples con e, mais, ou, pero. Oraciones temporales (cando) y comparativas y superlativas básicas. Afirmaciones, negativas, interrogativas con partículas y exclamativas básicas con partículas. Subordinadas simples con conjunciones más frecuentes: así que, porque, cando... Características de cohesión: Preposiciones y locuciones preposicionales más frecuentes. Uso de pronombres sólo con referente claro. Elipse sólo de elementos conocidos y muy claros. Marcadores de orden del discurso, de espacio y de tiempo más frecuentes (hoxe, agora, cedo, lonxe, preto, enriba...)

- Nivel 2: Incluye mensajes de carácter personal (SMS, correos, cartas) o de carácter social con frases rutinarias, cuestionarios sencillos, notas y mensajes relacionados con sus actividades de trabajo, estudio y ocio y textos sociales breves tipificados (para felicitar, invitar, aceptar/rehusar, agradecer, solicitar un servicio, pedir disculpas). Géneros textuales del nivel 1 y también páginas web, recetas, periódicos y revistas (titulares, noticias y anuncios) , horóscopos, cuentos y novelas cortos, anuncios de trabajo... Instrucciones sobre seguridad, aparatos de uso frecuente, alojamiento, menú de restaurante, vida académica, sanidad, posología de medicamentos... Información sencilla de formularios y documentos administrativos (Correos, Administración Pública, bancos, entidades académicas...). Registro formal e informal. Adverbios más comunes de cantidad (moito, pouco), tiempo (hoxe, agora, cedo), lugar (enriba, lonxe, preto), modo (ben, mellor, amodo), exclusión (só, soamente), inclusión (até, mesmo, incluso), afirmación, negación y duda. Características léxico-conceptuales: Nombres comunes y propios. Interjecciones más frecuentes. Vocabulario de situaciones cotidianas y actividades habituales (por especial interés) y un vocabulario receptivo más amplio que permita comprender textos (sentimientos, meteorología, fauna y flora, alimentación, vida académica, salud, Administración, tiempo libre, actividades profesionales más habituales). Sufijo- iño/a. Afijos comunes relativamente frecuentes. Superlativo absoluto y relativo. Comparativos analíticos y sintéticos más frecuentes, forma meirande. Usos de ti, vós / vostede, vostedes. Características verbales: Verbos copulativos. Verbos regulares y semirregulares (durmir, servir, espir) e irregulares más frecuentes (ser, estar, facer, saber, querer, ir, vir, haber, poder, ter, deber, etc.). Reflexivos y pronominales más comunes (chamarse, esquecerse, queixarse, lavarse, vestirse). Tiempos de indicativo en presente, pasado y futuro. Futuro hipotético. Infinitivo, gerundio y participio. Presente de subjuntivo para expresión de deseos, sentimientos y reacciones (Oxalá veñas!/Que teñas sorte!/Sinto que marches). Imperativo para consejos, instrucciones y órdenes. Perífrases más habituales: estar, andar + gerundio, volver, comezar a, ter que, estar a/para, estar a/andar a + infinitivo, ter + participio, haber (de) + infinitivo. Características de estructura sintáctica: Coordinadas copulativas (e, mais, nin), adversativas (mais, pero, senón) y disyuntivas (ou...ou, nin...nin). Oraciones impersonales

sencillas. Afirmaciones, negativas (incluyendo doble negación), interrogativas con partículas, exclamativas e imperativas. Subordinadas simples con conjunciones más frecuentes: así que, porque, cando, ademais... Subordinadas temporales (cando, ao + inf) , causales, finales, condicionales y comparativas (máis...ca, meirande...ca). Características de cohesión: Preposiciones, locuciones preposicionales, conectores y enlaces más frecuentes. Uso de prons sólo con referente claro. Elipse sólo de elementos conocidos y muy claros. Marcadores de orden del discurso, de espacio y de tiempo más frecuentes y también entón, logo, por outra banda...

- Nivel 3: Incluye textos breves en lengua estándar relacionados con la vida cotidiana que informen sobre acontecimientos o transmitan opiniones, deseos u órdenes. También textos sencillos sobre temas de su ámbito académico o profesional. Entre otros: cartas de restaurantes, formularios, anuncios, folletos, visitas turísticas, hoteles, alquiler, cartas y correos, relatos, anuncios o artículos de la prensa sobre temas de actualidad, prospectos de medicamentos, folletos de divulgación y publicitarios... Características léxico-conceptuales: Nombres comunes y propios. Vocabulario de situaciones cotidianas, actividades de ocio, sentimientos, del ámbito laboral y de temas de interés general, como salud, accidentes, tecnología, medio natural, economía, sociedad, geografía, servicios... Características verbales: Voz pasiva. Tiempos de indicativo en presente, pasado y futuro, futuro hipotético y presente de subjuntivo de los verbos regulares y de los irregulares más comunes. Imperativo (también negativo) para consejos, instrucciones y órdenes. Perífrasis verbales frecuentes: poder + infinitivo (obrigación) , deber (de) + infinitivo (obrigación) , haber + inf. (obrigación) , ter + part. (aspecto) , andar + gerundio, ir + gerundio, estar para + infinitivo, poñerse a + infinitivo), volver + infinitivo, empezar a + infinitivo, acabar de + infinitivo, deixar de + infinitivo, levar + gerundio, seguir + gerundio, dar + participio... Características de estructura sintáctica: Coordinadas copulativas (e, mais, nin) , adversativas (mais, pero, senón) y disyuntivas (ou...ou, nin...nin) , distributivas (ben... ben, ora... ora) y explicativas (é dicir, ou sexa, isto é). Usos de si: oraciones impersonales, pasivas y verbos reflexivos. Afirmaciones, negativas (incluyendo doble negación) , interrogativas con partículas (también indirectas) , exclamativas e imperativas. Subordinadas con subjuntivo e infinitivo. Subordinadas adjetivas y sustantivas, adverbiales (causales, finales, consecutivas, condicionales, temporales). Estilo indirecto. Características de cohesión: Preposiciones, locuciones preposicionales, conectores y enlaces más frecuentes. Uso de pronombres sólo con referente claro. Elipse sólo de elementos conocidos y muy claros. Marcadores de orden del discurso, de espacio y de tiempo más frecuentes y también entón, logo, por outra banda...

[Instrucción] Tienes que aprender los niveles definidos anteriormente y analizar los ejemplos que aparecen a continuación. Después, se te enviarán una serie de textos para que los clasifiques de uno en uno, por orden de aparición. Una aclaración: los textos contienen los caracteres “[SEP]”, que debes interpretar como saltos de línea en el texto. Cada texto aparece en una línea o fila. Estas son las tareas que debes realizar:

- Leer en profundidad el texto
- Utilizar las características de los niveles de complejidad descritos para analizar el texto, buscando, contando, señalando y extrayendo todas las que aparezcan en el texto
- Utilizar todas las características extraídas para determinar de qué nivel es el texto
- Responder con el nivel del texto basándote en los pasos anteriores, indicando solamente un número del 1 al 4

Estos pasos los tienes que repetir para cada texto.

[Contexto]

La información de los niveles con sus características para analizar aparecen en este mensaje, en la sección [Objetivo].

[Inputs esperados dentro de este mensaje]

- Niveles de complejidad y sus características
- Lista de textos para clasificar

[Formato y criterios de salida]

- Tabla csv con el nivel de cada texto de entrada, por orden.

[Restricciones]

- No inventes características que no estén en la descripción de características.
- No definas el nivel de un texto sin estar seguro de que es acorde.
- No respondas información que no sea el nivel del texto.

[Errores comunes a evitar]

- Inventar características
- No hacer el recuento de las características correctamente
- Inventar niveles

13.1. English translation of the prompt used in LLM experiments

[Role] You act as an expert linguist in Galician, specializing in classifying texts by complexity. You are an expert at classifying texts into the complexity levels defined below. You respond to the input texts that appear in the following message by clearly indicating which level they belong to.

[Objective] Your objective is to classify input texts according to their complexity into four levels: three defined levels and a fourth undefined level that includes texts more complex than those defined. Classifications must be clear and assigned to only one of the levels. The levels will be indicated only by a number from 1 to 4, where 1 corresponds to level 1, 2 to level 2, 3 to level 3, and 4 to level 4. You will be provided with a series of examples for each level to help you better understand the task. These are the characteristics of the different levels:

- Level 1: Includes everyday texts, such as short emails or messages related to daily activities, signs, lists, manuals, advertisements, forms, materials related to their interests, and brief authentic documents (tickets, admission tickets, restaurant menus, invoices, labels, maps, packaging, and schedules) in very frequent use, simple instructions, easy stories, song lyrics, and simple poems. Includes the most common adverbs of quantity (much, little), time (today, now, soon), place (above, far, near), manner (well, better, slowly), exclusion (only, solely), and affirmation and negation. Lexical-conceptual features: Common and proper nouns. Suffix -iño/a. Meaning of the most frequent affixes. Superlative -ísimo/a. Vocabulary limited to everyday situations and topics of special interest. Verbal features: Copulative verbs. Frequently used regular and semi-regular verbs (to sleep, to serve, to undress). Most common irregular verbs (to go, to do, to be). Most common reflexive and pronominal verbs (to be called, to forget, to complain). Indicative tenses in the present, past, and future. Hypothetical future. Most common periphrases: ir (to go) + infinitive, estar/andar (to be/to go) + gerund, voltar (to return) + infinitive, comezar a (to begin to) + infinitive, ter que (to have to) + infinitive. Syntactic structure features: Simple coordinating conjunctions such as “and,” “more,” “or,” and “but.” Temporal clauses (when) and basic comparative and superlative forms. Affirmative, negative, and interrogative sentences with particles, and basic exclamatory sentences with particles. Simple subordinate clauses with the most common conjunctions: so, because, when... Cohesion features: The most common prepositions and prepositional phrases. Use of pronouns only with a clear referent. Ellipsis only of known and very clear elements. The most common discourse, spatial, and temporal markers (today, now, soon, far, near, above...)
- Level 2: Includes personal messages (text messages, emails, letters) or social messages containing routine phrases, simple questionnaires, notes, and messages related to work, study, and leisure activities, as well as short, standard social texts (for congratulating, inviting, accepting/declining, thanking, requesting a service, and apologizing). Text types from Level 1, as well as web pages, recipes, newspapers, and magazines (headlines, news, and advertisements), horoscopes, short stories and novellas, job postings... Safety instructions, frequently used devices, housing, restaurant menus, academic life, healthcare, medication dosages... Simple information from forms and administrative documents (Post Office, Government, banks, academic institutions...). Formal and informal register. Most common adverbs of quantity (a lot, a little), time (today, now, soon), place (above, far, near), manner (well, better, slowly), exclusion (only, just), inclusion (up to, even, including), affirmation, negation, and doubt. Lexical-conceptual features: Common and proper nouns. Most frequent

interjections. Vocabulary for everyday situations and common activities (of special interest) and a broader receptive vocabulary that allows for understanding texts (feelings, weather, flora and fauna, food, academic life, health, government, leisure, most common professional activities). Suffix -iño/a. Relatively frequent common affixes. Absolute and relative superlatives. Most frequent analytical and synthetic comparatives, "meirande" (greater) form. Uses of "ti, vós / vostede, vostedes (you). Verb characteristics: Copulative verbs. Regular and semi-regular verbs (to sleep, to serve, to undress) and the most common irregular verbs (to be, to exist, to do, to know, to want, to go, to come, to have, to be able to, to have, to must, etc.). Most common reflexive and pronominal verbs (to be called, to forget, to complain, to wash, to get dressed). Indicative tenses in the present, past, and future. Hypothetical future. Infinitive, gerund, and participle. Present subjunctive for expressing wishes, feelings, and reactions (I hope you come!/Good luck!/I'm sorry you're leaving). Imperative for advice, instructions, and commands. Most common periphrases: estar/andar (to be/to go) + gerund, volver a (to return) + infinitive, empezar a (to begin to) + infinitive, ter que (to have to) + infinitive, estar a punto de (to be about to) + infinitive, ir a (to be going to) + infinitive, haber + participle, haber de (to have to) + infinitive. Syntactic structure features: Coordinating conjunctions (and, more, nor), adversative conjunctions (but, however, instead), and disjunctive conjunctions (either...or, neither...nor). Simple impersonal sentences. Affirmative, negative (including double negation), interrogative with particles, exclamatory, and imperative sentences. Simple subordinate clauses with the most common conjunctions: so, because, when, also... Temporal subordinate clauses (when, as + infinitive), causal, final, conditional, and comparative (more...than, "meirande" (greater)... than). Cohesion features: Prepositions, prepositional phrases, the most common connectors and linking words. Use of pronouns only with a clear referent. Ellipsis only of known and very clear elements. The most common discourse, spatial, and temporal markers, as well as "then," "later," "on the other hand," etc.

- Level 3: Includes short texts in standard language related to everyday life that report on events or convey opinions, requests, or instructions. Also includes simple texts on topics within their academic or professional field. Among others: restaurant menus, forms, advertisements, brochures, tourist information, hotels, rentals, letters and mail, stories, news reports or press articles on current events, medication leaflets, informational and advertising brochures... Lexical-conceptual features: Common and proper nouns. Vocabulary related to everyday situations, leisure activities, feelings, the workplace, and topics of general interest, such as health, accidents, technology, the natural environment, the economy, society, geography, and services... Verbal features: Passive voice. Present, past, and future indicative tenses, future subjunctive, and present subjunctive of regular verbs and the most common irregular verbs. Imperative (including negative) for advice, instructions, and commands. Common verbal periphrases: poder (can/may) + infinitive, deber (de) (must/should) + infinitive, haber de (to have to) + infinitive, ter (to have to) + participle, andar (to go) + gerund, ir (to go) + gerund, estar para (to be about to) + infinitive, poñerse a (to start/begin) + infinitive, volver a (to do again) + infinitive, empezar a (to begin to) + infinitive, acabar de (to have just) + infinitive, dejar de (to stop) + infinitive, levar (to have been) + gerund, seguir (to continue) + gerund, dar (to give) + participle. Syntactic features: Coordinating conjunctions (and, more, nor), adversative conjunctions (more, but, rather), disjunctive conjunctions (either...or, neither...nor), distributive conjunctions (either...or, or...or), and explanatory conjunctions (that is, in other words, that is to say). Uses of "si": impersonal, passive, and reflexive verb clauses. Affirmative, negative (including double negation), interrogative with particles (including indirect), exclamatory, and imperative. Subordinate clauses with subjunctive and infinitive. Adjectival and noun clauses, adverbial clauses (causal, final, consecutive, conditional, temporal). Indirect speech. Cohesion features: Prepositions, prepositional phrases, the most frequent connectors and linking words. Use of pronouns only with a clear referent. Ellipsis only of known and very clear elements. The most frequent discourse, spatial, and temporal markers, as well as then, later, on the other hand...

[Instructions] You must familiarize yourself with the levels defined above and analyze the examples provided below. You will then be sent a series of texts to classify one by one, in the order they appear. Note: The texts contain the characters "[SEP]", which you should interpret as line breaks in the text. Each text appears on a single line. These are the tasks you must perform:

- Read the text thoroughly
- Use the characteristics of the complexity levels described to analyze the text, searching for, counting, identifying, and extracting all those that appear in the text
- Use all the extracted characteristics to determine the text's level

- Answer with the text's level based on the previous steps, indicating only a number from 1 to 4

You must repeat these steps for each text.

[Context]

The information on the levels and their characteristics to be analyzed appears in this message, in the [Objective] section.

[Expected inputs within this message]

- Complexity levels and their characteristics
- List of texts to classify

[Output format and criteria]

- CSV table with the level of each input text, in order.

[Restrictions]

- Do not invent features that are not listed in the feature description.
- Do not assign a level to a text unless you are certain it is appropriate.
- Do not provide information other than the text's level.

[Common Mistakes to Avoid]

- Inventing characteristics
- Not counting characteristics correctly
- Inventing levels

13.2. Examples added at the end of the prompt for few-shot experiments:

[Ejemplos]

[Texto][Nivel]

['-Cal é o teu horario de traballo? [SEP] -Depende. Algúns días traballo moitas horas e outros case non traballo. O que si, sempre me ergo tarde, ás dez ou ás once. [SEP] Despois, almorzo e logo dou un paseo. Ao mediodía volvo á casa, xanto algo e vexo a tele un pouco. [SEP] Logo, baixo á rúa e collo o coche. [SEP] Ás veces traballo ata as nove ou as dez da noite. Cando acabo, vou á casa, preparo a cea e leo un pouco. Déitome á unha ou ás dúas máis ou menos. [SEP] Xaquín Pereira, 38 anos, taxista.'][1]

['Cabalgata e Recepción dos Reis Magos [SEP] venres, 5 xaneiro, 2024 - 17:00 [SEP] Praza do Obradoiro, [SEP] Santiago de Compostela [SEP] Cabalgata e Recepción dos Reis Magos [SEP] Os Reis Magos de Oriente percorrerán as rúas da cidade ata a praza do Obradoiro. [SEP] Ao rematar terá lugar a tradicional recepción no Pazo de Raxoi para todos os nenos e nenas que queiran achegarse ás súas Maxestades. [SEP] PERCORRIDO PREVISTO: [SEP] Saída ás 17:00 horas desde a Praza da Mercé e desfile pola avenida de Ferrol, Frei Rosendo Salvado, praza Roxa, República de El Salvador, rúa do Hórreo, praza de Galicia, rúa da Senra, Porta Faxeira, rúa do Vilar, praza das Praterías, travesa de Fonseca e praza do Obradoiro. A recepción está prevista para as 19.00 horas.'][1]

['Con esta receita saen uns 12 Sapos da Limia xeitosos. Se queredes que vos saian máis, só tedes que multiplicar os ingredientes. E non vos preocupedes polo sabor a cervexa, que lles queda de vicio! [SEP] Ingredientes [SEP] 85 ml de cervexa [SEP] 125 ml de leite [SEP] 1 ovo [SEP] 3 culleradas de azucre [SEP] 1 belisco de sal [SEP] 230 gramos de fariña triga [SEP] 20 gramos de manteiga derretida [SEP] Aceite de xirasol para fritir [SEP] Azucre para rebozar sapos [SEP] Preparación [SEP] Imos mesturando os ingredientes por orde de aparición na listaxe. [SEP] Deixamos que repouse a masa media hora fóra da neveira. [SEP] Quentamos o aceite de xirasol nun caciño e imos botando culleradas de masa. Aquí a dica é botar toda a masa de golpe. Eu fíxeno coa culler dos xeados. [SEP] Rebozamos os sapos en azucre e xa estarían!'][1]

['As festas de Ribeira [SEP] Este ano as festas en honra a santa Uxía contarán coa presenza das dúas orquestras de referencia no ámbito musical galego. A orquestra Panorama e a París de Noia animarán o

día central das festas de Ribeira, desde as primeiras horas da noite ata ben entrada a madrugada. Non só está garantida a calidade musical, senón tamén o espectáculo, a diversión e as ganas de bailar ata moi, moi tarde. Ademais, a comisión de festas agasallará a todos os asistentes coa tradicional queimada ao finalizar a verbena. Para que a diversión non decaia, agosto comeza en Ribeira!'][1]

['SANTIAGO DE COMPOSTELA [SEP] TAPAS [SEP] Para tomar tapas, o máis recomendable é facelo na zona vella, nas rúas do [SEP] Franco e da Raíña, onde están todos os bares con especialidades galegas: polbo á feira, empanada, pementos de [SEP] Padrón, mariscos... Algúns dos máis famosos son O Orella, onde podes tomar orella de porco, o SantYago, [SEP] Os Caracois, o Central, todos eles na rúa da Raíña. Son bares moi animados onde se mestura moita xente nova, estudantes, peregrinos e turistas. [SEP] ONDE COMER [SEP] A Barrola, con comida galega de boa calidade e a bo prezo; o Enxebre, ao lado da catedral; El Pasaje, con especialidade en mariscos e carnes; [SEP] A Alameda, con especialidade en empanada; a Casa Marcelo, comida galega moderna, ten unha estrela [SEP] Michelin.'][1]

['Do Alto do Acevo a Castroverde [SEP] Distancia: 45,1 km [SEP] Nos seus primeiros metros, o Camiño Primitivo atravesa unha fraga con carballos, bidueiros e castiñeiros. Esta primeira etapa cruza unha contorna de montaña, na que se coroan diferentes altos, como o da Fontaneira ou o da Vacariza. Nos vales, os prados e pastizais engaden textura á paisaxe. E entre a vexetación cómpre sinalar os piñeirais, tanto da especie do país como as repoboacións de piñeiro silvestre. [SEP] 1. Alto do Acevo [SEP] 43°8'49.3"N 6°58'24"W [SEP] O Acevo, a máis de mil metros de altitude, recibe o nome polos acivros que hai nesta zona. O gando doméstico que pastorea nos montes emprega as acivreiras para refuxiarse do frío no inverno e da calor no verán. Tamén acubillan aves como o paporroibo, o gaio ou o merlo común. [SEP] 2. Conxunto de acivros [SEP] 43°8'36.8"N 6°58'9.9"W [SEP] O acivro é un arbusto con follas perennes de cor verde escura. É resistente ao frío e medra devagar. Florece durante a primavera e os seus froitos, mantenza da fauna silvestre, maduran no seguinte inverno, presentando unha cor vermella e brillante. A súa madeira é dura e resistente. [SEP] 3. Alto da Fontaneira [SEP] 43°2'15.2"N 7°11'45.5"W [SEP] A súa altitude, a 936 metros, e a súa situación no interior de Galicia, fan que o clima deste lugar sexa de montaña. Caracterízase por ter unha temperatura media baixa, polo que son frecuentes as xeadas. Son abundantes tamén as precipitacións, en forma de neve no inverno. [SEP] 4. Fraga de Estornín [SEP] 43°1'44.11"N 7°12'32.96"W [SEP] As fragas son extensións de monte, xeralmente de difícil acceso, poboadas por árbores caducifolias. A de Estornín ten carballos, pradairos, capudres, teixos, bidueiros e castiñeiros. No seu interior habitan mamíferos salvaxes abundantes na zona, como lobos, corzos, xabarís ou raposos.'][2]

['Detida unha parella na Coruña que se fixo pasar por policía para roubar un móbil e unha carteira [SEP] A Policía Nacional detivo a un home e a unha muller que se fixeron pasar por policías para roubar un teléfono móbil e unha carteira na Coruña, ademais doutros roubos. Segundo informou o Instituto Armado, o primeiro dos roubos ocorreu na zona de Oza, no mes de agosto, cando abordaron a unha persoa que camiñaba soa pola rúa. [SEP] A muller, esgrimindo unha navalla, esixiulle que lle entregase as súas pertenzas. Así, conseguiron roubarlle o teléfono móbil e a súa carteira, que contiña diñeiro en efectivo e cartóns bancarios. Días máis tarde, ao redor das 7.00 horas da mañá, cometeron outro roubo na rúa Costa Rica cando a vítima ía coller un taxi e é sorprendida pola parella. [SEP] A parella fíxose pasar por policías e argumentaron que realizaban un control de drogas. Con esta escusa, empuxaron á vítima contra unha parede e subtraéronlle o seu teléfono móbil e a súa carteira para, posteriormente, abandonar o lugar. Nos dous casos, nas horas posteriores aos roubos, realizaron varios cargos cos cartóns bancarios subtraídos en diferentes establecementos da cidade.'][2]

['Falece o xornalista Pepe Seijo [SEP] Seijo, de 57 anos, estaba traballando nos informativos cando faleceu de maneira repentina. [SEP] Unha das voces máis características das ondas radiofónicas en Lugo, Pepe Seijo, faleceu esta segunda feira de forma repentina mentres traballaba no informativo diario de Radio Lugo-Cadena SER, segundo informa a TVG. O xornalista, de 57 anos de idade, entrou en directo no primeiro boletín da xornada, mais no segundo xa non interveu. [SEP] Seijo naceu en Bilbao e pasou a súa infancia en Euskadi. Seguidor apaixonado do Athletic de Bilbao, como se recoñecía el mesmo en calquera conversa, chegou a ser presidente da peña luguesa deste equipo. [SEP] Décadas de traballo ligado ás ondas servíronlle para converterse nunha voz de referencia na cidade e arredores, que perdeu no principio do ano outro dos seus xornalistas radiofónicos máis coñecidos, Arcadio Silvosa.'][2]

['Vigo prenderá as luces de Nadal o 24 de novembro ao que suma unha árbore de 44 metros, "a máis importante do mundo" [SEP] O alcalde de Vigo, Abel Caballero sinalou ese día como o momento en que acenderá "o Nadal de todo o planeta". [SEP] O alcalde de Vigo, Abel Caballero, confirmou este xoves que o acto de aceso da iluminación do Nadal terá lugar finalmente o 24 de novembro ás 20,30

horas, momento en que se acenderá "o Nadal de todo o planeta". [SEP] Así o trasladou en declaracións aos medios nunha visita á Porta do Sol, onde este xoves comezaron os traballos de instalación do gran abeto luminoso que, este ano, alcanzará os 44 metros de altura. "A árbore máis importante do mundo é o de Vigo", presumiu o rexedor, que explicou que a estrutura contará cunha gran estrela na súa cúspide, cunha lonxitude total de 19 metros.'][2]

['Os cinco preceptos nas ensinanzas de Buda [SEP] Como toda relixión, o budismo conta con preceptos básicos que deben seguirse con rectitude. En total, só hai cinco, pero abarcan áreas importantes da vida. Os preceptos de Buda son "Non mates", "Non roubes", "Non abuses do sexo" e "Non consumas drogas nin alcohol". Entende a continuación a razón de cada un. [SEP] Non mates [SEP] É posible que toda relixión, filosofía ou doutrina teña en conta esta lei. As ensinanzas de Buda van un pouco máis aló que outras tradicións, porque cando di non mates -porque formas parte do todo e cometendo tal acto estás a facerte dano- tamén está a falar de animais, como a galiña, o boi ou ata unha formiga. [SEP] Non roubes [SEP] Se non queres o que pertence aos demais e estás satisfeito cos teus logros, xa vas por bo camiño. Pero aínda así, o budismo subliña a idea de que non se debe roubar, aínda que sexa o lugar de alguén na fila, froito do esforzo intelectual ou físico de alguén, ou mesmo de obxectos. [SEP] Non abuses do sexo [SEP] O sexo é absolutamente natural e moi ben visto no budismo, porén non deixa de ser un intercambio de enerxía e calquera exceso é visto de forma atenta polas ensinanzas de Buda. Polo tanto, é importante manter o acto sexual saudable e como complemento da túa vida, non como foco das relacións. [SEP] Non Consumas Drogas nin Alcol [SEP] Mantén a túa mente activa e sempre en plenitude, observar o momento presente é fundamental para lograr chegar a Magga, é dicir, a fin do sufrimento. Por outra banda, o uso de estupefacientes -legalizados ou non- altera o funcionamento do cerebro e, polo tanto, o seu uso non se recomenda no budismo.'][3]

['O CLIENTE, A NOSA PRIORIDADE [SEP] Sempre ao seu servizo [SEP] O noso obxectivo é ofrecer aos clientes a mellor calidade aos mellores prezos, cun bo servizo e atención en centros de proximidade. Ademais, contamos cunha ampla variedade en frescos, produtos de marcas líderes e marca Froiz. O cliente está no centro de todas as decisións: "Sempre ao seu servizo", tentando dar resposta ás súas demandas e necesidades. [SEP] FROIZ [SEP] Nós [SEP] Somos unha empresa dedicada a venda por xunto e polo miúdo de produtos frescos, alimentación, adega e droguería, a través da nosa rede comercial formada por supermercados, hipermercados, cashcarry e tenda en liña. [SEP] A nosa actividade desenvólvese en España e Portugal. En España, estamos presente nas comunidades autónomas de Galicia, Castela e León, Castela-A Mancha e Comunidad de Madrid. A empresa sitúase entre as 20 primeiras empresas do sector a nivel nacional e forma parte da central de compras Euromadi. [SEP] COMPROMETIDOS COS NOSOS VALORES [SEP] O noso obxectivo empresarial é ofrecer aos clientes a mellor calidade aos mellores prezos, cun bo servizo e atención en centros de proximidade. [SEP] CALIDADE [SEP] OS MELLORES PREZOS [SEP] SERVIZO E ATENCIÓN [SEP] FORNECIDO AMPLO [SEP] PROXIMIDADE E COMODIDADE'][3]

['Artigo 1. Obxecto e finalidade [SEP] 1. Esta orde ten por obxecto fixar as bases reguladoras do Programa para a promoción do emprego autónomo en Galicia (TR341D) e proceder á súa convocatoria para o ano 2024. [SEP] 2. A finalidade deste programa é a concesión de 2 liñas de axuda económica a aquelas persoas desempregadas que pretendan desenvolver a súa actividade empresarial ou profesional en Galicia como traballadoras autónomas ou por conta propia, para facer fronte aos distintos gastos xerados no comezo e mantemento da súa actividade laboral. [SEP] 3. Ao abeiro desta orde subvencionaranse as altas na Seguridade Social ou en mutualidade de colexio profesional e os gastos de mantemento da actividade, que, cumprindo os requisitos e condicións establecidas nela, se formalicen desde o 30 de setembro de 2023 ata o 29 de setembro de 2024, ambos inclusive. [SEP] Establécense 2 liñas de axudas: [SEP] a) Liña 1: unha axuda para o inicio da actividade económica e mantemento do emprego como persoa traballadora autónoma. [SEP] b) Liña 2: unha axuda para sufragar os gastos do mantemento da actividade por conta propia, equivalente a 12 meses da contía da cota reducida regulada no apartado 1 do artigo 38.ter da Lei 20/2007, do 11 de xullo, do Estatuto do traballo autónomo. [SEP] Nesta convocatoria incorpórase o establecemento de métodos de custos simplificados conforme ao disposto no Regulamento (UE) 2021/1060 do Parlamento Europeo e do Consello, do 24 de xuño de 2021.'][3]

['OS HEAVIES E AS POETAS [SEP] Ao remate dun recital, un xornalista espétame: "Os heavies e os poetas sodes os últimos inocentes deste mundo". Nunca tal eu pensara, e iso que sempre me prestaron moito os heavies. Inspírame confianza a xente que responde á dureza das cousas armándose de dureza, desde o nome até o coiro. Teñen a sensibilidade dos que precisan responder ás agresións.

Teñen conciencia grupal. Non son submisos. Para os demais, cadeas e reloxos. Eles non van á moda, por iso teñen aínda o corazón de ouro. Os poetas, en cambio, habiamos levar máis ferro ao corazón. [SEP] Parada como quedei despois daquela arroutada, non fun quen de responder que non debería haber cousa menos inocente que un poeta. O profesor Antón Figueroa, referíndose ao campo cultural, fala adoito do "mito da inocencia". Os mitos dan lugar aos prexuízos. Os prexuízos impiden coñecer. Alguén di "leliadoura", e o lector, con bo tino, conclúe "plátano es". Os poetas adoitan deixar as frases en suspenso e os poemas, xa se sabe, son cousa moi ambigua. O plátano vén das Canarias. En Canarias é unha hora menos. En Galicia perdemos luz por culpa do cambio horario. (Isto non é poesía.) [SEP] Conclusión: a cultura come e dá de comer en boa medida grazas á implantación social do mito da inocencia. Que boa xente, os artistas. E os poetas, eses xa son o máximo. [SEP] Escriben con metáforas. Eles, menudo estilo. E elas, tan estilosas. [SEP] Hai moitos xeitos de perder a inocencia. Todos son bos. Algúns mecanismos de lexitimación son tan visíbeis que case non é preciso denuncialos. En cambio, adóitase ignorar a presenza efectiva (e por sutil, se cadra, dobremente efectiva) de modos de intervención máis inmediatos. Como se escribir publicamente non fose xa un xeito de exercer o poder. [SEP] A linguaxe tamén é violencia. Sempre que se abre un turno de palabra, outra persoa deixa de falar. A alternativa, quizais: non escribir para que nos escoiten, senón para que nos respondan. Non creo que o poder sexa malo de seu, e Foucault demostrou que vén de abaixo, hipótese que convida a exercelo até arriba. Non somos inocentes, as poetas, e non é mala cousa. A vida é complicada, e a vida cultural non é unha excepción. [SEP] Todos temos intereses, pero non todos os intereses son iguais. Quizais cumprise comezar por distinguir entre os públicos e os privados. Entre os persoais e os colectivos. [SEP] Para poder quedar, hai que asumir que estamos. Que non se nos xulgue pola presenza ou ausencia de estratexias, senón polo sentido das nosas estratexias. O que facemos ou deixamos de facer non será cabal na medida en que careza de propósito, senón na medida en que o seu propósito sexa máis ou menos cabal. [SEP] Deámonos o luxo de ser obxectados publicamente polo que procuramos, e non só polo que facemos e dicimos. As cousas moitas veces non son o que parecen. Para poder saír do reino da opinión, habería que comezar por dicir, canda Montaigne, "o que eu opino non dá a medida do mundo, senón a medida do meu entendemento". O mundo é ancho e alleo. Do entendemento, en poucos casos podemos dicir que é propio. Fagamos o posíbel por que non sexa estreito.’][4]

Plan-Guided Text Simplification with Extended Contexts

Pascal Mathas, Jan Bakker, Jaap Kamps
Institute for Logic, Language and Computation (ILLC)
University of Amsterdam
Amsterdam, The Netherlands
pascal.mathas@student.uva.nl, j.bakker@uva.nl, kamps@uva.nl

Abstract

In this paper, we investigate the impact of increasing context lengths (one to five paragraphs) on plan-following accuracy in plan-guided text simplification. Plan-guided models simplify text according to sentence-level operation labels such as copy, rephrase, split, and delete. Previous work fine-tunes BART with target reading-level and sentence-level operation tokens to perform this task. We find that BART's plan-following accuracy on Newsela-auto drops significantly as context increases from one to five paragraphs. This means that the model becomes less reliable with longer contexts, and the quality of its outputs decreases. To address this, we propose replacing the fine-tuned BART models with a prompting-based approach using instruction-tuned Qwen models. We find that this approach not only maintains robust plan-following across all context lengths, but even at the longest context length still exceeds BART's performance at the shortest. We further provide ablations on model size and model family, showing that a minimum model capacity is required for the approach to work and that it transfers across LLM families.

Keywords: text simplification, plan-guided generation, long-context, large language models, BART

1. Introduction

Plan-guided simplification models, as proposed by Cripwell et al. (2023b), simplify text according to a plan. This plan consists of target reading levels and sentence-based operation labels. The model receives in its input the target reading level token, plus sentence-level operation tokens for each sentence. These tokens include <*copy*>, <*rephrase*>, <*split*>, and <*delete*>, and are inferred from references (oracle) or by a separate planning model designed to predict the appropriate token for each sentence (Cripwell et al., 2023b).

In this work, we show that fine-tuned plan-guided BART simplification models, introduced by Cripwell et al. (2023b), degrade in their plan-following ability as the context length increases. Since document simplification requires broader context to preserve discourse structure and handle multi-sentence operations (Alva-Manchego et al., 2019; Cripwell et al., 2023a), this degradation therefore exposes a limitation in the fine-tuned BART models. We investigate whether this gap can be addressed by replacing the fine-tuned BART models with prompting instruction-tuned large language models (LLMs).

The main contributions of our paper are as follows:

- We show that fine-tuned plan-guided BART models degrade in their plan-following accuracy as context increases.
- We show that a prompting-based approach using instruction-tuned LLMs maintains both plan-following accuracy and simplification quality at extended context lengths.

We make all code, prompts, and settings publicly available at github.com/pascalmathas/plan_simp_extended.

2. Related Work

Sentence alignment. Jiang et al. (2020) propose a neural Conditional Random Field (CRF) alignment model that leverages the sequential structure of sentences in parallel documents, combined with fine-tuned BERT to capture semantic similarity. The model aligns simple sentences to complex sentences, after which simplification operation labels can be inferred from the alignments (copy, rephrase, split, delete). In their work, the authors introduce the Newsela-auto and Wiki-auto datasets. In our work, we use the Newsela-auto dataset to train and evaluate our models, and the neural CRF aligner to align our generations back to the complex sentences.

Plan-guided simplification. The baseline BART planning model we use in this paper is adapted from Cripwell et al. (2023b). The authors introduce two models: a planning model that predicts a sequence of sentence-level operations and a generation model conditioned on these plans. They show that their plan-guided system outperforms other end-to-end approaches. In Cripwell et al. (2023a), the authors extend this work by using broader contextual information during generation. In this work, we show that plan-guided BART models degrade in plan-following ability as input context increases. Due to some missing details in the pipeline of Cripwell et al. (2023b), we follow the code from a re-

Proceedings of the Joint Workshop on Readability and Text Simplification (READIxTSAR) @ LREC 2026, pages 121–129
11 May 2026.

production of those works by Bakker and Kamps (2024).

Prompting strategies for LLM-based simplification. Finally, to address the degradation observed in BART models, we replace them with LLMs. Our approach is loosely inspired by Papandreou et al. (2025), where the authors investigate several prompting strategies. We adapt their codebase as a starting point for our LLM implementation, although our prompting strategy and focus differ.

3. Task & Experimental Setup

3.1. Data

The data used for our experiments is the Newsela-auto dataset from Jiang et al. (2020), based on the Newsela corpus by Xu et al. (2015). Each of the $1,882$ news articles in the corpus is manually rewritten at five different simplification levels. Our preprocessing approach follows that of Cripwell et al. (2023b). Since this paper only requires complex-to-simple article pairs, we pair each article version with every version corresponding to a simpler reading level, resulting in a total of $18,820$ article pairs. We split the data into training, validation, and test sets of 92.5%, 2.5%, and 5.0%, respectively.

Table 1 shows the dataset statistics of Newsela-auto after preprocessing, where $|c_i|$ is the average token length of a complex sentence, $|s_i|$ is the average token length of a simple sentence, n and k are the average number of sentences in the complex and simple documents, and p is the average number of sentences per paragraph.

	Newsela-auto
# Doc Pairs	18,820
# Para Pairs	478,479
# Sent Pairs	960,365
Avg. $\|c_i\|$	17.14
Avg. $\|s_i\|$	12.84
Avg. n	51.03
Avg. k	43.25
Avg. p	2.01

Table 1: Statistics of the Newsela-auto dataset after preprocessing, where n is # sentences in C, and k is # sentences in S and p is # sentences per paragraph in C.

Paragraphs in the Newsela-auto dataset are relatively short, containing roughly two sentences on average. This underlines the importance of evaluating on multiple input paragraphs, as real-world text is likely to contain longer paragraphs.

We infer the operation labels based on the already aligned sentences in the Newsela-auto dataset. A complex sentence is labeled as *delete* if it has no aligned counterpart, *split* if it aligns to multiple simplified sentences, and *rephrase* if it aligns to a single, different sentence. Sentences with a Levenshtein similarity of ≥ 0.92 to their aligned counterpart are labeled as *copy*. Table 2 shows the distribution of operation labels in the Newsela corpus. Our distribution differs slightly from that of Cripwell et al. (2023b), as we did not have access to their filtered version and thus used the unfiltered variant.

Data	**Copy**	**Rephrase**	**Split**	**Delete**
Newsela-auto	20.14	27.21	16.69	35.95

Table 2: Operation class distributions of Newsela-auto in percentages.

Finally, to create the paragraph-level datasets for $i = 1 \ldots 5$, each document in the data splits is divided into chunks of i paragraphs. If a document is not evenly divisible, chunks of size $i-1$ are used to distribute the remainder evenly. For example, if a document contains 7 paragraphs, then for $i = 3$, the document is chunked into paragraph groups of 3–2–2.

3.2. Planning Models

To assess how the models would perform in a real-world scenario, we also evaluate them with predicted labels instead of oracle labels. For this, we use the liamcripwell/pgdyn-plan model checkpoint made available on Huggingface, which was part of Cripwell et al. (2023a) and trained on the oracle labels. As the model checkpoint is the context-aware variant of the planning model, we first generate context representations of the test documents. We then use these contexts when predicting operation labels for the test set.

3.3. Evaluation

We evaluate the simplifications produced by our models in two ways. First, we assess simplification quality at the document level. Second, we evaluate the models' ability to follow the operation labels when simplifying sentences.

3.3.1. Alignment

To align the generated simplifications with the complex sentences, we use the neural CRF alignment model of Jiang et al. (2020). Instead of training the aligner ourselves, we use a Wiki-auto pretrained checkpoint from Bakker and Kamps (2024). The checkpoint is available on GitHub: aligner checkpoint. We align the full outputs to the inputs instead

of aligning each paragraph pair separately. This means that we group the paragraphs back into the full documents using their `pair_id`.

3.3.2. Document-Level Metrics

After generating simplifications with the models, we evaluate them by regrouping the paragraphs into documents and calculating several document-level metrics. We use two reference-based metrics: (a) SMART, which measures semantic similarity between the output and reference using sentence embeddings (Amplayo et al., 2023), and (b) SARI, which assesses simplification quality by comparing n-gram operations (additions, deletions, and unchanged) between the output, source, and reference (Xu et al., 2016). Additionally, we use the reference-free Flesch-Kincaid Grade Level (FKGL), which measures text readability (Kincaid et al., 1975).

3.3.3. Plan-Following (Micro-Recall)

To assess whether the models follow the operator labels, we first align the generated simplifications back to the complex sentences. We then assign operator labels using the same procedure described in Section 3.1. Once the labels for the generated simplifications are obtained, we compare them to the reference labels to calculate the percentage of sentences that were simplified according to the oracle plan. We refer to this metric as micro-recall, the percentage of sentences simplified according to the oracle plan.

3.4. Computational Requirements

Table 3 presents the computational requirements for our experiments. The reported time estimates correspond to all five runs together, that is, for all paragraph-level datasets $i = 1 \ldots 5$. Micro-recall time includes generation, alignment, and micro-recall computation. Generating with the LLMs is significantly faster due to the vLLM implementation.

Model	GPU	Training	Micro-recall	Doc-eval
BART (all variants)	A100	∼6h	∼12h	∼2h
Qwen-7B & Qwen-14B	A100	–	∼9h	∼2h
Qwen-32B	H100	–	∼9h	∼2h

Table 3: Computational requirements for training and inference.

3.5. Reproducibility

All fine-tuning experiments in this paper are run with a fixed random seed (42) and deterministic Torch and CUDA behavior to ensure reproducibility.

For a detailed overview of all parameters for every run, refer to the *README* and *shell scripts* in our GitHub repository.

4. Problem: BART Model Variants Degrade in Plan-Following as Context Increases

4.1. Baseline Model Setup

We train a total of 10 BART models: 5 with oracle labels (plan-guided) and 5 without. Both groups are trained on paragraph-level datasets with $i = 1 \ldots 5$. We refer to the BART models with oracle labels as O-BART$_{\text{para-}i = 1 \ldots 5}$ and the models without oracle labels as BART$_{\text{para-}i = 1 \ldots 5}$.

Both groups follow the same training procedure. The base model we fine-tune is facebook/bart-base (Lewis et al., 2019). Training is performed using the Adam optimizer with a learning rate of 2×10^{-5} and an effective batch size of 16 (batch size 8 with gradient accumulation of 2). We use mixed precision (FP16) and early stopping with a patience of 1. For both models, target reading level tokens are prepended to each paragraph, along with operation tokens for O-BART$_{\text{para}}$ following Cripwell et al. (2023b). During inference, we use beam search with a beam size of 5.

We also evaluate the performance of the BART models with predicted operation labels from the planning model. For this, we use the BART models trained with oracle labels, but during inference, we use the predicted labels. We refer to these models as $\hat{\text{O}}$-BART$_{\text{para-}i = 1 \ldots 5}$.

4.2. Baseline Analysis

Figure 1 showcases the micro-recall for the BART model variants across paragraph input lengths of 1 to 5.

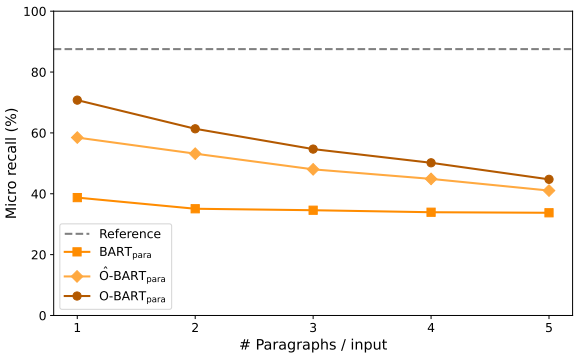


Figure 1: Plan-following micro recall for BART models across context sizes on Newsela-auto.

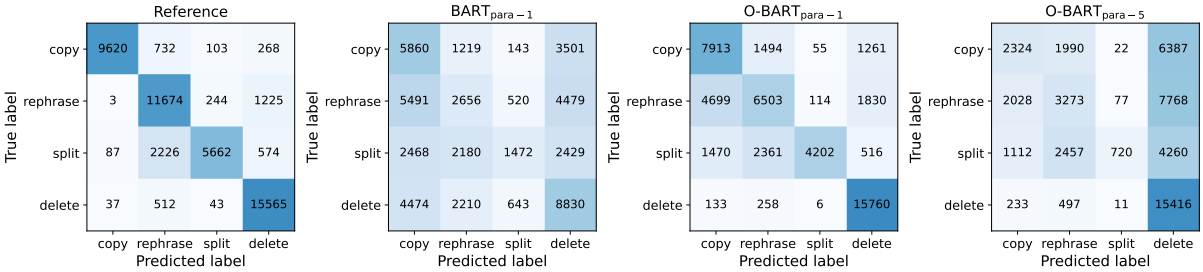


Figure 2: Confusion matrices comparing the ground-truth (oracle) operation labels against the operation labels predicted by the model for input documents, evaluated on the Newsela-auto dataset.

O-BART achieves the highest plan-following accuracy at paragraph level 1 (70.8%) but degrades as context increases, dropping to 44.7% at paragraph level 5, a loss of 26.1 percentage points. Ô-BART follows a similar decline, starting at 58.5% and falling to 41.0%, being limited by the planning model. The unguided BART baseline remains relatively flat (38.7% to 33.8%), showing that without plans the model defaults to a standard set of operations. Interestingly, at paragraph level 5, all models are relatively close to each other, indicating that increased context degrades the models' performance to almost baseline levels.

This becomes more apparent when we look at the matrices in Figure 2, where we compare the oracle labels (y-axis) to the labels predicted by the model (x-axis). For paragraph level 1, O-BART follows the copy and delete operations somewhat reliably but already struggles with rephrase and split. At paragraph level 5, copy recall drops from 73.8% to 30.9%, rephrase from 49.5% to 33.1%, and split from 49.2% to 13.7%. Nearly half of all copy and rephrase sentences are incorrectly deleted, indicating that as context grows, the model increasingly defaults to deletion rather than executing the intended operation.

In Table 4, we can observe the document-level simplification results for the BART model variants.

System	SMART ↑			FKGL ↓	SARI ↑	Length	
	P	R	F1			Tok.	Sent.
Input	58.9	64.9	61.5	7.61	19.5	1089.9	49.1
Reference	100	100	100	4.62	100	721.1	45.2
BART$_{para-1}$	57.7	53.3	55.1	6.15	38.7	712.1	37.8
O-BART$_{para-1}$	61.5	60.6	60.8	6.02	45.3	807.7	40.7
Ô-BART$_{para-2}$	60.8	54.6	57.3	6.96	44.6	847.9	33.0
Ô-BART$_{para-3}$	58.8	50.9	54.4	7.93	43.3	946.7	29.2
Ô-BART$_{para-4}$	59.7	45.8	51.6	8.25	42.3	781.6	23.5
Ô-BART$_{para-5}$	59.1	40.7	47.9	9.17	40.7	667.9	18.8
O-BART$_{para-1}$	66.1	61.0	63.4	5.93	50.3	709.8	36.5
O-BART$_{para-2}$	65.4	55.0	59.7	6.89	48.4	744.7	29.7
O-BART$_{para-3}$	63.0	51.3	56.5	7.80	47.1	826.0	26.3
O-BART$_{para-4}$	63.8	46.2	53.4	8.18	45.2	689.3	21.2
O-BART$_{para-5}$	62.9	41.0	49.4	8.99	42.6	588.5	17.0

Table 4: Results of document simplification of BART models on Newsela-auto.

We can see that the operation labels help the model produce better simplifications, as the Ô-BART models achieve higher scores than base BART across the board. The delete bias at higher paragraph levels, however, can also be seen here, as the number of output sentences drops drastically. FKGL worsens with increased paragraph input for both variants, surpassing even the input FKGL. As the model produces fewer sentences, these sentences are more complicated and longer than the input.

5. Solution: Prompting Instruction-tuned LLMs

5.1. Model & Inference

To explore whether prompting instruction-tuned LLMs results in better plan-following accuracy at extended context sizes, we replace the fine-tuned BART model with a prompting-based approach. For the prompting-based approach, we use Qwen2.5-7B-Instruct, Qwen2.5-14B-Instruct, and Qwen2.5-32B-Instruct (Team, 2024; Yang et al., 2024). Our setup is inspired by Papandreou et al. (2025), but we introduce several modifications. First, we utilize the vLLM (Kwon et al., 2023) library for increased inference speed and memory efficiency. Second, we adjust the prompt so the LLM knows to follow the four operation labels when simplifying. All models are run in bfloat16, and for generation we set the temperature to 0.2, top-p to 0.9, repetition penalty to 1.1, and max new tokens to 2048.

As a small ablation to assess the generalizability of our prompting-based pipeline to a different LLM family, we also run google/Gemma-3-27B-Instruct (Gemma-27B) (Team, 2025) at paragraph levels 1 and 5, using the same parameters as the Qwen models.

5.2. Prompting Strategies

Similar to the BART models, the LLM receives in each user prompt the target reading level to simplify to and, before each sentence, the operation labels indicating what it should do with that sentence. In the system prompt, we first indicate that this is a simplification task and that it should simplify according to the operation labels. We then explain the reading levels and operation labels. Next, we introduce several rules, for example, that <*rephrase*> should differ from its input, and that <*copy*> should produce exactly the same output. Finally, we provide three few-shot examples of the training data.

We also run Qwen-32B without access to oracle labels. For this variant, we adjust the prompt to not refer to the operation labels and remove them from the examples, while keeping the reading levels. This prompt serves as our unguided baseline for the prompting-based approach.

The two prompts that we used can be found in Appendix A and on our GitHub: *prompts*.

6. Results

Figure 3 shows the micro-recall results of the fine-tuned BART models compared to the prompting-based approach with the Qwen model variants.

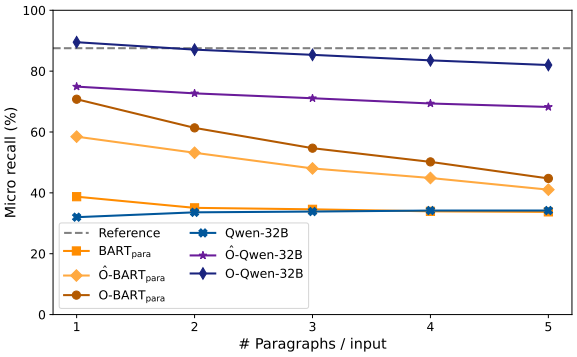


Figure 3: Plan-following micro recall for BART and Qwen-32B models across context sizes on Newsela-auto.

O-Qwen-32B achieves 89.5% micro-recall at paragraph level 1 and drops to 82.0% at 5 input paragraphs, a loss of 7.5 percentage points compared to 26.1 for O-BART over the same range. Ô-Qwen-32B similarly declines less steeply. Notably, Ô-Qwen-32B (with predicted operation labels) at 5 input paragraphs still exceeds O-BART (with oracle labels) at 1 paragraph, suggesting that the prompting-based approach is more robust to increased context even under imperfect planning. The unguided LLM performs similarly to the unguided BART model.

Input (Complex)

<copy> Those competing interests were already playing out on fire-stripped slopes. <split> Above, the woodpeckers were feasting on wood-boring beetles that began swarming dead trees while they were still smoking. <rephrase> Below, crews with chain saws and big-rig trucks were removing barriers and salvaging fallen tree trunks from roads and paths created for power lines. <copy> “We’re looking for silver linings,” Bridgman said with a sigh. <copy> “But we’re caught between extremes.”

O-BART$_{para\text{-}5}$ Simplification

Those competing interests were already playing out on fire-stripped slopes. Above, the woodpeckers were feasting on wood-boring beetles that began swarming dead trees while they were still smoking.

O-Qwen-32B-5 Simplification

Those competing interests were already playing out on fire-stripped slopes. Above, the woodpeckers were feasting on wood-boring beetles. They began swarming dead trees while they were still smoking. Below, crews with chain saws and big-rig trucks were removing barriers. They were also salvaging fallen tree trunks from roads and paths created for power lines. “We’re looking for silver linings,” Bridgman said with a sigh. “But we’re caught between extremes.”

Figure 4: Simplification comparison between O-BART$_{para\text{-}5}$ and O-Qwen-32B-5. Operation labels are marked by <>.

The ability of Qwen to follow the plan much more faithfully also becomes apparent when we look at the matrices in Figure 5, where O-Qwen-32B-1 follows the plan with high accuracy, and O-Qwen-32B-5 still maintains strong performance compared to the deletion bias of O-BART. This deletion bias is illustrated in Figure 4: O-BART deletes the final three sentences (rephrase, copy, copy) and fails to split, while O-Qwen-32B correctly executes all operations.

System	SMART ↑			FKGL ↓	SARI ↑	Length	
	P	R	F1			Tok.	Sent.
Input	58.9	64.9	61.5	7.61	19.5	1089.9	49.1
Reference	100	100	100	4.62	100	721.1	45.2
Qwen-32B	39.0	43.9	41.2	5.01	40.0	923.3	53.4
Ô-Qwen-32B-1	57.5	61.4	59.2	5.51	47.4	830.5	48.9
Ô-Qwen-32B-2	57.3	60.5	58.7	5.54	47.3	815.9	47.9
Ô-Qwen-32B-3	57.2	59.5	58.2	5.57	47.2	794.4	46.7
Ô-Qwen-32B-4	57.3	58.5	57.7	5.63	47.0	777.0	45.2
Ô-Qwen-32B-5	57.2	57.8	57.3	5.68	47.0	766.6	44.3
O-Qwen-32B-1	61.9	62.0	61.9	5.49	51.5	743.3	43.8
O-Qwen-32B-2	61.7	61.2	61.4	5.53	51.2	729.9	42.9
O-Qwen-32B-3	61.8	60.3	61.0	5.59	51.0	712.6	41.9
O-Qwen-32B-4	62.0	59.3	60.5	5.62	50.9	693.0	40.5
O-Qwen-32B-5	61.8	58.6	60.1	5.68	50.6	684.5	39.7

Table 5: Results of document simplification of LLM models on Newsela-auto.

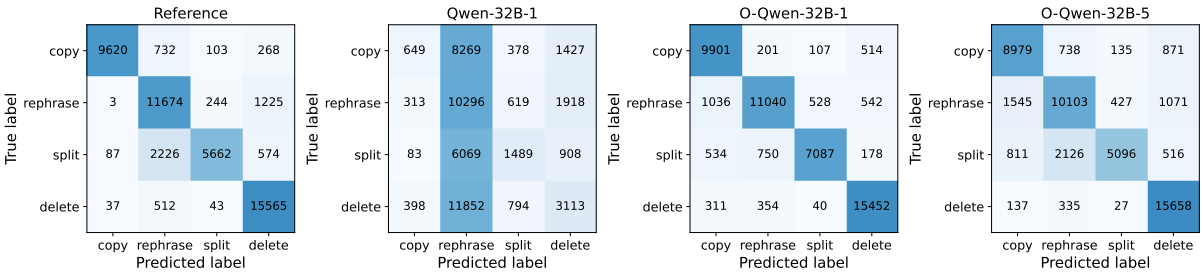


Figure 5: Confusion matrices comparing the ground-truth (oracle) operation labels against the operation labels predicted by the model for input documents, evaluated on the Newsela-auto dataset.

Finally, the document-level results in Table 5 show that both O-Qwen-32B and Ô-Qwen-32B remain stable in SMART, FKGL, and SARI scores across input levels, with only marginal drops. The Qwen-32B model with predicted operation labels performs only slightly below the model with oracle labels. Ô-Qwen-32B-5 produces more sentences than O-Qwen-32B, staying closer to the reference length. Finally, even though Qwen-32B achieves the lowest FKGL among the generated models (approaching the reference), it is semantically worse, as indicated by its lower SMART and SARI scores.

6.1. Effect of Model Size

Figure 6 shows the effect of model size on plan-following in the prompting-based approach.

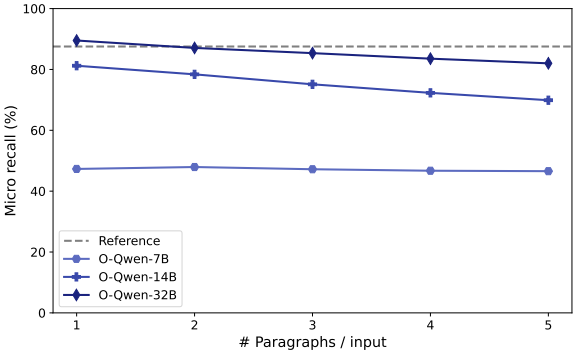


Figure 6: Plan-following micro recall for O-Qwen-7B, O-Qwen-14B, and O-Qwen-32B models across context sizes on Newsela-auto.

O-Qwen-7B is unable to execute the plan for any given input length, performing comparably to the unguided baselines. O-Qwen-14B achieves reasonable plan-following at paragraph level 1 (81.2%) but degrades to 69.9% at 5-paragraph inputs, a drop of 11.3 percentage points. O-Qwen-7B-5 achieves similar performance to the unguided baseline (SARI 40.5 vs. 40.0), indicating that the 7B model lacks the capacity to accurately simplify the input according to the operation labels.

System	SMART ↑			FKGL ↓	SARI ↑	Length	
	P	R	F1			Tok.	Sent.
Input	58.9	64.9	61.5	7.61	19.5	1089.9	49.1
Reference	100	100	100	4.62	100	721.1	45.2
Qwen-32B	39.0	43.9	41.2	5.01	40.0	923.3	53.4
O-Qwen-7B-5	49.6	54.3	51.5	6.15	40.5	834.1	47.0
O-Qwen-14B-1	59.5	59.7	59.5	5.80	49.9	749.5	43.4
O-Qwen-14B-2	59.8	58.3	59.0	6.01	49.4	726.0	40.6
O-Qwen-14B-3	59.4	56.6	57.9	6.13	48.8	702.3	38.8
O-Qwen-14B-4	59.1	54.8	56.8	6.30	48.2	678.7	36.8
O-Qwen-14B-5	58.9	53.7	56.1	6.42	47.5	662.3	35.7
O-Qwen-32B-5	61.8	58.6	60.1	5.68	50.6	684.5	39.7

Table 6: Results of document simplification of O-Qwen models on Newsela-auto.

O-Qwen-14B-1 shows comparable performance to O-BART$_{\text{para}}$ (Table 6), but degrades less as context grows. This is again reflected in the number of output sentences at 5 input paragraphs compared to O-BART$_{\text{para-5}}$ (35.7 vs. 17.0), indicating that the prompting-based approach avoids the deletion bias even at smaller model sizes. O-Qwen-32B maintains the strongest performance across the board.

6.2. Effect of Model Family

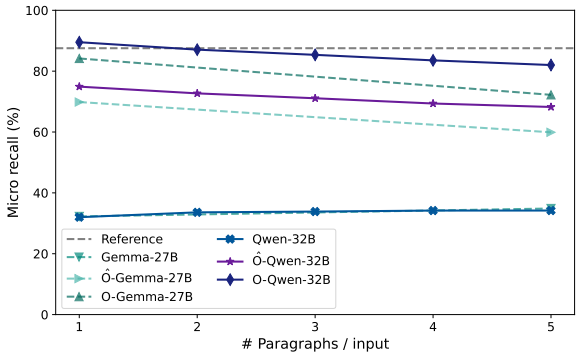


Figure 7: Plan-following micro recall for Qwen-32B and Gemma-27B model variants across context sizes on Newsela-auto.

Figure 7 and Table 7 show the results of running the same prompting-based pipeline with Gemma-27B. O-Gemma-27B-1 achieves 84.2% micro-recall, dropping to 72.2% at 5 input paragraphs, a loss of 12.0 percentage points. This is larger than O-Qwen-32B (-7.5) but well below O-BART (-26.1). Ô-Gemma-27B follows a similar pattern, declining from 69.8% to 59.9%. The document-level metrics in Table 7 show that O-Gemma-27B performs below O-Qwen-32B across SMART and SARI, and is less robust than the Qwen model family when context increases.

System	SMART ↑			FKGL ↓	SARI ↑	Length	
	P	R	F1			Tok.	Sent.
Input	58.9	64.9	61.5	7.61	19.5	1089.9	49.1
Reference	100	100	100	4.62	100	721.1	45.2
Gemma-27B	36.0	40.3	37.9	4.75	38.7	955.7	53.3
O-Gemma-27B-1	52.5	55.5	53.8	4.97	48.5	778.7	47.2
Ô-Gemma-27B-5	49.7	49.2	49.3	5.00	45.8	695.0	41.3
O-Gemma-27B-1	56.7	56.0	56.3	4.99	51.5	688.4	41.9
O-Gemma-27B-5	53.6	49.8	51.5	5.01	47.9	615.7	36.7

Table 7: Results of document simplification of Gemma-27B models on Newsela-auto.

7. Discussion

Fine-tuning vs. prompting. As shown in Section 6, the prompting-based approach achieves stronger plan-following performance than the fine-tuning approach at every paragraph input level, while also degrading less as context increases.

The degradation in the performance of the fine-tuned BART model is likely due to the difficulty of learning the sentence-level operation token semantics through fine-tuning. As more paragraphs are included, the number of operation tokens in the input grows, making the association between tokens and their corresponding sentences increasingly difficult to learn. At $i = 1$, BART sees ~ 2 operation tokens on average. At $i = 5$, it sees ~ 10. The relatively small capacity of BART ($\sim$140M parameters) likely amplifies this difficulty.

The instruction-tuned Qwen models do not have this degradation for multiple reasons. First, the meaning of each operation token is defined in the prompt rather than being learned through fine-tuning, and the models are specifically trained to follow structured instructions. Second, operation tokens are naturally interleaved with their corresponding sentences in the prompt, making the association between each token and its sentence straightforward regardless of input length. Besides this, the Qwen models are also significantly larger, the main model of this paper being at 32B parameters (vs. 140M).

Model size. BART and Qwen-32B have a substantial difference in model capacity. However, our ablation of different model sizes in Section 6.1 shows that the two approaches have very different capacity requirements. O-BART outperforms O-Qwen-7B, a model 50 times its size, demonstrating that fine-tuning can be effective at smaller scales. The prompting-based approach requires more capacity, with O-Qwen-14B being the minimum size at which the model can effectively follow operation labels and O-Qwen-32B further improving performance.

Model family. As shown in Section 6.2, the prompting-based approach transfers to Gemma-27B, though with lower plan-following and simplification quality than Qwen-32B. This gap may partly be explained by the difference in model size (27B vs. 32B), and by the fact that our pipeline was developed and tested using Qwen. It is therefore possible that Gemma-27B could perform better with, for instance, tweaked generation parameters or prompt adjustments.

Computational costs. The prompting-based approach requires more computational resources at inference time, primarily due to the larger model sizes (we used an H100 80GB for Qwen-32B). This cost is, however, somewhat mitigated by the non-existent training time and fast generation with vLLM. Additionally, O-Qwen-14B achieves comparable plan-following to O-BART while degrading far less at extended contexts, and requires only an A100 40GB.

7.1. Limitations

Dataset. The Newsela-auto dataset is proprietary. To gain access to the data, we first had to request access to the Newsela corpus from Newsela, and then contact the authors of Jiang et al. (2020) to obtain the Newsela-auto dataset. This significantly hinders reproducibility and further research.

Scope. Another limitation of our work is that we evaluate the prompting-based approach on only one dataset, which limits generalizability to other domains. Additionally, while we include Gemma-3-27B as an ablation, our main findings are based on a single LLM family (Qwen), and further validation across a broader range of models and datasets would strengthen our conclusions.

8. Conclusion

In this paper, we have shown that fine-tuned plan-guided BART models degrade in their plan-following ability as context grows. To address this, we have replaced the fine-tuned BART models with a prompting-based approach using instruction-tuned LLMs. We have shown that this approach

maintains both plan-following accuracy and simplification quality at extended context lengths, degrading far less as input context increases. Even with predicted operation labels from the planning model, the prompting-based approach at 5 input paragraphs exceeds the fine-tuned BART model with oracle labels at 1 paragraph.

Lay Summary

We look at how well different AI systems follow detailed editing plans when asked to simplify longer pieces of text. The system is given sentence-by-sentence instructions, such as "shorten this" or "keep this." Earlier models stop following those instructions once the input grows from one paragraph to several, even after extensive training. Instead, they tend to delete too much or make things up, which hurts the quality of the simplified text. The researchers test a different approach using large language models guided through prompting rather than fine-tuning. This approach handles longer texts much better: it sticks to the requested edits even when given several paragraphs at once, and produces higher-quality simplifications in every setting.

Acknowledgments

Jan Bakker and Jaap Kamps are supported by the Netherlands Organization for Scientific Research (NWO NWA # 1518.22.105). Jaap Kamps is also supported by the University of Amsterdam (AI4FinTech program) and ICAI (AI for Open Government Lab). Views expressed in this paper are not necessarily shared or endorsed by those funding the research.

Bibliographical References

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2019. Cross-sentence transformations in text simplification. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 181–184.

Reinald Kim Amplayo, Peter J. Liu, Yao Zhao, and Shashi Narayan. 2023. SMART: sentences as basic units for text evaluation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Jan Bakker and Jaap Kamps. 2024. Beyond sentence-level text simplification: Reproducibility study of context-aware document simplification. In *Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context@ LREC-COLING 2024*, pages 27–38.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2023a. Context-aware document simplification. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13190–13206.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2023b. Document-level planning for text simplification. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 993–1006.

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. 2020. Neural crf model for sentence alignment in text simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7943–7960.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Taiki Papandreou, Jan Bakker, and Jaap Kamps. 2025. Medical text simplification from jargon detection to jargon-aware prompting. In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility and Readability (TSAR 2025)*, pages 36–46.

Gemma Team. 2025. Gemma 3.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

A. Prompts

A.1. Oracle Prompt

The full oracle prompt includes three few-shot examples, omitted here for brevity. The prompt with the examples can be found on our GitHub: *oracle prompt*.

```
You are a text simplification editor.  Each input sentence is numbered and
labeled with an operation.  Execute each operation and output the result with
matching sentence numbers.  Output in English only.

Reading levels indicate the target simplification level:  <RL_0> (most complex)
to <RL_4> (simplest).  Higher reading levels should produce simpler output for
REPHRASE and SPLIT operations.

Operations:
- COPY: Output the sentence with ZERO changes.  Do not fix, improve, or edit
anything.
- REPHRASE: You MUST modify the sentence.  Remove or simplify at least one
phrase, clause, or difficult word.  The result must differ from the input
while preserving the core meaning.
- SPLIT: Break into exactly 2 shorter sentences at a natural point.  Keep as
many original words as possible.
- DELETE: Delete this sentence entirely.  Do not include its number in your
output.

Output format:
- Write one numbered line per non-deleted sentence, matching input numbers.
- For SPLIT, write both result sentences on the same numbered line.
- For DELETE, skip that number.
- Output ONLY the numbered results.  No explanations.
```

A.2. Unguided Prompt

The full unguided prompt includes three few-shot examples, omitted here for brevity. The prompt with the examples can be found on our GitHub: *unguided prompt*.

```
You are a text simplification
editor.  Simplify the given text.
Output in English only.

Reading levels indicate the target
simplification level:  <RL_0> (most
complex) to <RL_4> (simplest).
Higher reading levels should
produce simpler output.
```

LLM-Generated Stories for Students with Significant Cognitive Disabilities: Promise, Gaps, and Evaluation Framework

Pragati Maheshwary, Ananya Ganesh, Shamya Karumbaiah
University of Wisconsin-Madison
Madison, WI, USA
{pmaheshwary, aganesh27, shamya.karunbaiah}@wisc.edu

Abstract

Students with significant cognitive disabilities (SCD) require specially designed accessible stories for reading comprehension assessments, yet creating such content is labor-intensive and difficult to scale. This preliminary study investigates whether large language models (LLMs) can generate short accessible stories for alternate assessment system. Using an 8-fold cross-validation design, we generated 120 stories with GPT-4o via one-shot prompting with human-written exemplars and evaluated them against a test set comprising 7 expert-human written stories as baselines across three dimensions: simplicity, fluency & coherence, and thematic adherence. Cross-validation results show that generated stories meet surface-level simplicity targets, with approximately two-thirds falling within the human baseline range for readability metrics. However, generated stories exhibited a systematic coherence gap where only 5% fell within the human range for adjacent sentence similarity, a pattern consistent across all folds. Thematic adherence was moderate, with adequate diversity across stories. These findings suggest LLMs can serve as a drafting tool within accessible content generation pipelines, but human expert review remains essential to ensure coherence, testability, and alignment with quality standards required for high-stakes alternate assessments.

Keywords: accessible story generation, text simplification, cognitive disabilities, reference-free evaluation

1. Introduction

Students with significant cognitive disabilities (SCD) represent a heterogeneous population that includes learners with intellectual disability, autism, multiple disabilities and more, who have complex communication and language needs (Karvonen and Clark, 2019; Thurlow et al., 2016). Under the Individuals with Disabilities Education Act (IDEA) and the Every Student Succeeds Act (ESSA), states are required to include these students in accountability systems through alternate assessments based on alternate achievement standards (AA-AAS), with participation capped at approximately 1% of the total tested population (Thurlow et al., 2017). The Dynamic Learning Maps (DLM) alternate assessment system is one of the largest operational AA-AAS programs in the United States, serving approximately 90,000 students with SCD across more than 18 states (Karvonen and Clark, 2019; Karvonen et al., 2021). Within the DLM system, English Language Arts (ELA) assessments are designed for students to engage with short accessible stories and then respond to reading comprehension items aligned with *Essential Elements* (the alternate content standards that guide instruction and assessment for this population).

Creating accessible high-quality stories for this population is a significant challenge. Each story must adhere to accessibility criteria such as the use of simple vocabulary drawn from high-frequency word lists, short sentences with explicit referents, clear narrative structures that adhere to a singular theme, and minimal inference load and ambiguity. These criteria are grounded in what is known about the communication and language profiles of students with SCD, a substantial proportion of whom communicate primarily using one or two words, signs, or symbols at a time (Nash et al., 2016), and many of whom rely on augmentative and alternative communication (AAC) devices (Erickson and Geist, 2016). As a result, human-authored assessments undergo an extensive multi-stage development pipeline that includes initial drafting by trained item writers, peer review, multiple rounds of internal quality control by content and accessibility specialists, external review for content accuracy and bias & sensitivity, editorial review, and field testing before stories become operational (DLM Consortium, 2024). This process, while essential for ensuring quality and validity, is labor-intensive, time-consuming, and difficult to scale to meet the growing demand for diverse, engaging, and grade-appropriate content.

This exploratory study attempts to introduce LLMs as drafting tools that adhere to baseline quality controls to fast-track the multi-stage development pipeline for such stories by utilizing existing advancements in the field of text simplification and AI-assisted content generation. We investigate two primary questions: (1) How do LLMs perform at the task of short story generation that is inspired by existing literary sources but adapted for students with SCD? (2) How do LLM-generated stories for SCDs compare with expert-human written stories on established metrics from text simplification re-

Proceedings of the Joint Workshop on Readability and Text Simplification (READIxTSAR) @ LREC 2026, pages 130–141
11 May 2026.

search? To answer these questions, we employ a cross-validation experimental design in which LLM-generated stories and human-written baselines are both scored using a multi-dimensional evaluation framework. The evaluation framework assesses stories across three key dimensions, (1) simplicity, (2) Fluency & Coherence, and (3) Thematic Adherence. This study makes two primary contributions. First, we introduce a reference-free evaluation framework with the aforementioned criteria for assessing LLM-generated accessible stories that matter for this specialized context. Second, we make a case for where LLMs as a drafting tool can fit in the pipeline for generating accessible stories using one-shot prompting with human-written exemplars, highlighting both their promise and gaps. Together, these contributions lay the groundwork for responsibly leveraging AI to supplement – not replace – human expertise in creating accessible educational content and assessment for students with significant cognitive disabilities.

2. Related Work

The broader field of text simplification has long sought to make written content more accessible to diverse reader populations. Traditional approaches relied on rule-based systems involving lexical substitution and syntactic simplification, as well as statistical methods that treated simplification as a form of monolingual translation using parallel corpora such as Wikipedia and Newsela (Xu et al., 2016). More recently, neural approaches, including sequence-to-sequence models and transformer-based architectures such as BERT, T5, and GPT variants, have substantially advanced the state of the art in text simplification (Alva-Manchego et al., 2020; Martin et al., 2022). These models have been evaluated using both traditional readability formulas such as Flesch-Kincaid Grade Level and Flesch Reading Ease, as well as task-specific metrics including SARI (Xu et al., 2016) for measuring simplification quality and BLEU and ROUGE for content preservation. Benchmarks such as ASSET (Alva-Manchego et al., 2020) and TurkCorpus have further supported systematic evaluation. Notably, Chamovitz and Abend (2022) demonstrated that incorporating cognitively motivated simplification operations, such as reducing syntactic complexity and resolving ambiguous references, can improve the quality of simplified text beyond what surface-level transformations achieve. At the document level, Vásquez-Rodríguez et al. (2023) have highlighted the importance of maintaining coherence when simplifying longer texts, showing that simplification must attend not only to sentence-level readability but also to the logical flow and connectedness of the overall text. However, the vast majority of this work has targeted general adult readers, second language learners, or individuals with low literacy levels. Very little simplification research has explored the needs of students with SCDs, whose reading and communication profiles differ from these other populations and require linguistic simplification and structural clarity that extend well beyond what standard simplification approaches typically produce (Yalon-Chamovitz, 2009).

Large language models (LLMs) have recently demonstrated strong performance across a wide range of natural language generation tasks, including creative writing, summarization, and text simplification through zero-shot and few-shot prompting (Brown et al., 2020; OpenAI, 2023). In the educational domain, there has been growing interest in using LLMs to support assessment development. Tan et al. (2025) provided a comprehensive review of automatic item generation techniques leveraging LLMs, documenting the rapid evolution of these approaches and noting both their promise for producing diverse item pools and the persistent challenges related to quality assurance and alignment with assessment specifications. Laverghetta Jr. and Licato (2023) showed that LLMs can generate items for cognitive assessments that approximate the psychometric properties of human-written items, and further developed a framework for using LLMs to generate and validate psychometric items, demonstrating the feasibility of integrating AI into the assessment development cycle. Beyond item generation, LLMs have also been explored for narrative content creation. Feng et al. (2025) developed a framework for generating social stories using LLMs, demonstrating their capacity to produce structured narratives that adhere to specific pedagogical conventions. Raffloer and Green (2025) investigated reader perceptions of AI-generated versus human-authored narratives, finding that while readers can sometimes detect differences, AI-generated stories can achieve comparable levels of narrative engagement under certain conditions. Despite these important advances in both AI-assisted item generation and narrative generation, the potential of LLMs to assist in the specific and labor-intensive task of drafting accessible stories for students with SCD has not yet been investigated.

3. Methods

This study employed a cross-validation experimental design to evaluate the quality of large language model (LLM)-generated accessible stories for students with cognitive disabilities. The design systematically compared AI-generated stories against human-written baseline stories using a quantitative evaluation framework inspired from some of the Dynamic Learning Maps (DLM) assessment criteria for accessible educational text.

3.1. Data

We manually extracted the reading comprehension stories written and evaluated by human experts from publicly available and retired DLM stories for Grade 11-12 English Language Arts [1]. These stories took inspiration from existing literary sources that are considered grade level appropriate readings recommend by State and Federal Educational Agencies. For the exploratory analysis in the current study, we focused only on generating and evaluating the story text and not on the questions/test items associated with the story for comprehension testing purposes. We ended up with 8 stories inspired from 3 source books, *The Great Gatsby, My Antonia and A White Heron*. Table 1 provides a detailed description about the stories associated with each book.

Table 1: Human Written Stories for Grade 11-12

Source Book	Story Title	Sentence Count	Word Count
My Antonia (MA)	Jim & Antonia	28	288
	Post Office	31	351
	The Garden	16	99
A White Heron (WH)	Mary & Martha	16	104
	The Businessman	22	129
	How about a Wig?	22	150
The Great Gatsby (GG)	The Valley of Ashes	20	195
	Nick Changes His Mind	26	157

3.2. Evaluation Framework

We developed a reference-free quantitative evaluation framework comprising three criteria partly aligned with DLM guidelines for accessible educational content (Table 2): Simplicity, Fluency & Coherence, and Thematic Adherence. The framework is reference-free by design for two reasons. First, the intended use case is to evaluate LLMs as drafting tools at the earliest stage of the novel content development pipeline, where a reference text does not exist. Second, because multiple distinct stories can be generated from the same source book, there is no one-to-one correspondence between human-written and LLM-generated texts that would make reference-based comparison meaningful (Belem et al., 2025).

We also adopted an automated quantitative framework as a necessary first step before human expert validation because collecting human judgments from accessibility specialists is time-consuming and expensive, and committing expert reviewer time is difficult to justify without first establishing that generated stories show sufficient baseline quality on well-validated metrics to warrant further review. The metrics selected for this framework are individually well-established in the text simplification and coherence literature (Kincaid et al., 1975; Vásquez-Rodríguez et al., 2023; Reimers and Gurevych, 2019). We note that these automated metrics do not capture all dimensions of quality that matter for DLM texts such as use of people-first language or fairness criteria that makes certain stories emphasizing biking over traveling in a car as unfair for students with motor disabilities. Therefore, human expert validation remains an essential next step in the broader scheme. The present framework is intended to provide a scalable, reproducible preliminary assessment that can identify systematic gaps before expert review resources are invested. The following subsections describe the operationalization of each criterion through the different metrics.

3.2.1. Simplicity

Readability was assessed using the Flesch-Kincaid Grade Level (FKGL; Kincaid et al., 1975) as the primary metric, computed as:

$$\text{FKGL} = (0.39 \times \text{ASL}) + (11.8 \times \text{ASW}) - 15.59 \quad (1)$$

where ASL = average sentence length (words) and ASW = average syllables per word. The Flesch Reading Ease (FRE) score was retained as a supplementary measure:

$$\text{FRE} = 206.835 - (1.015 \times \text{ASL}) - (84.6 \times \text{ASW}) \quad (2)$$

Average words per sentence and average syllables per word were also reported as interpretable diagnostics, as they are the direct inputs to both formulas.

3.2.2. Fluency and Coherence

Narrative coherence was assessed using three complementary measures. Entity continuity was computed as the proportion of adjacent sentence pairs sharing at least one proper noun (identified by capitalization) or personal pronoun, adapted from the entity-grid approach of (Vásquez-Rodríguez et al., 2023). Higher continuity scores indicate more referentially coherent narrative chains. Adjacent sentence similarity was measured as the mean TF-IDF cosine similarity between consecutive sentence pairs, capturing local topical flow. Sentence length standard deviation was used as a structural regularity indicator where high variance in sentence length beyond 5-12 words per sentence signals inconsistent adherence to accessibility constraints.

3.2.3. Thematic Adherence

Thematic adherence was operationalized with three different methodological approaches.

[1] Stories can be found at: https://monarchreader.com/home

Table 2: Evaluation Criteria, Metrics, and DLM Alignment

Dimensions	Metric	Interpretation Range	Purpose	DLM Alignment
Simplicity	Flesch-Kincaid Grade Level	Grade 0–16+ (lower = simpler)	Evaluates surface accessibility of syntax and vocabulary; lower grade level indicates simpler, more accessible text	Accessible Text Language: minimize inference load; Accessible Text Content: reduced depth, breadth, and complexity
	Flesch Reading Ease	0–100 (higher = easier)		
	Avg. words per sentence	Word count		
	Avg. syllables per word	Syllable count		
Fluency & Coherence	Entity Continuity Score	0–1 (higher = more coherent)	Assesses narrative logical flow, referential consistency, and structural regularity relative to DLM format constraints	Accessible Text Language: maintains logical structure and predictable sentence patterns
	Adjacent Sentence TF-IDF Similarity	0–1 (higher = smoother)		
	Sentence Length Standard Deviation	0–∞ (lower = more uniform)		
Thematic Adherence	Theme-Text Semantic Similarity	0–1 (higher = more coherent)	Measures whether each story's text semantically reflects its declared theme (embedding-based). Detects thematic collapse within a book's story set. Detects cross-book theme bleed within a fold	Instructional Relevance: preserves thematic intent and construct validity
	Within-Book Pairwise TF-IDF	0–1 (lower = more diverse)		
	Within-Fold Pairwise TF-IDF	0–1 (lower = more diverse)		

First, the generation prompt was designed to elicit an explicit theme statement from the LLM for each generated story, structured as "This [story/information text] explores [theme]." This approach was adopted because the source books (e.g., *My Antonia*) contain multiple themes across hundreds of pages, making direct comparison between a 15-30 sentence generated story and the full source text uninformative as a similarity signal. To assess whether each story's text reflect its declared theme, *single-story thematic consistency* was measured using embedding-based semantic similarity between the declared theme statement and the story text, computed as the cosine similarity between sentence embeddings produced by the `all-MiniLM-L6-v2` model (Reimers and Gurevych, 2019). Embedding-based similarity was preferred over lexical overlap measures because theme statements and story texts are expected to express the same meaning through different vocabulary – a story about perseverance will use words like "kept trying" and "finally succeeded" rather than the word "perseverance" itself, which lexical measures would penalize incorrectly.

Thematic diversity among the five stories generated for the same book within each fold was assessed by computing mean pairwise TF-IDF cosine similarity between all theme statements generated for the same target book within a fold. Low mean similarity indicates the LLM explored diverse themes naturally; high similarity indicates thematic collapse toward a dominant theme. This serves as both a quality indicator and a research finding about LLM behavior under one-shot prompting.

Within-fold thematic diversity was assessed to identify if themes overlap across source books within a fold, by computing mean pairwise TF-IDF cosine similarity across all 15 theme statements generated in a single fold, regardless of target book. If the LLM defaulted to the same thematic territory regardless of which book it was generating for, this would appear as high within-fold similarity. We chose TF-IDF cosine because comparison between theme statements concern lexical similarity — short, structurally parallel strings — rather than semantic coherence between a theme and a full story text.

Thematic adherence metrics were not computed for human-written stories, which were produced without LLM-declared theme statements. Human stories serve as a quality baseline for Simplicity, and Fluency & Coherence, but the Thematic Adherence criterion is specific to LLM-generated outputs.

3.3. LLM Text Generation

Stories were generated using OpenAI's GPT-4o model via the OpenAI API with temperature $= 0.4$

and $\texttt{max\_tokens} = 8000$. Each generation call used one-shot prompting, providing a single human-written text piece as an exemplar. The prompt instructed the model to generate five independent stories inspired by a specified target book, constrained to 15-30 sentences, using simple vocabulary and sentences, with characters and settings relating to the themes from a [target book]. The prompt additionally required the model to produce an explicit theme statement for each story which was used for thematic adherence evaluation as described in Section 3.2.3 (also see Appendix A).

3.4. Analysis

3.4.1. Baseline Characterization

Human-written texts were evaluated first to establish baseline distributions for all applicable metrics (i.e., Simplicity, and Fluency & Coherence). Descriptive statistics (means, standard deviations) were computed overall and by source book.

3.4.2. Cross-Validation Analysis

We implemented a multi-fold cross-validation design in which each human-written text served as the one-shot training exemplar exactly once. For each story fold:

- **Training set** ($n = 1$): One human text served as the exemplar in the generation prompt.
- **Test set** ($n = 7$): The remaining seven human stories served as quality baselines.
- **Generation**: The LLM produced five stories for each of the three source books, yielding 15 generated stories per fold.

This design yielded 120 total generated stories (8 folds $\times$ 3 books $\times$ 5 stories) while systematically varying the training exemplar to assess the influence of exemplar writing style and source book on generation quality.

For each of the 8 folds, the 15 generated stories were compared against the metric distributions of the 7 test stories in that fold. For each metric, we computed the proportion of generated stories whose value fell within the minimum–maximum range of the test stories. Because the training exemplar was excluded from the test set in each fold, this comparison is not contaminated by the story used to prompt the LLM. We then averaged these proportions across all folds to obtain a cross-validated estimate of how frequently generated stories achieve human-like metric values. We also computed the mean signed delta between each generated story's metric value and the test set mean to characterize the direction of any deviations. The proportion of generated stories falling within the test set range is reported overall, by training exemplar (at the fold level), by target book, and by prompting condition (same-source vs. cross-source). Mean signed deltas are reported at the overall level only. Thematic adherence is reported descriptively for generated texts only, as human stories were produced without LLM-declared theme statements.

4. Results

4.1. Human-Written Story Baseline Characteristics

Table 3 presents descriptive statistics for the eight human-written stories across the two evaluation criteria. These stories served as both the quality reference baseline and the one-shot training exemplars in the cross-validation design. Table 4 presents the same statistics aggregated by source book.

Simplicity. Human stories varied considerably in readability. FK Grade Level ranged from 1.97 (The Business Man) to 5.18 (Jim and Antonia), with a mean of 3.67 (SD = 1.04). This range is notable: even among texts approved for DLM Grade 11-12 administration, readability spans more than three grade levels, suggesting that the DLM accessible text standard accommodates substantial surface-level complexity variation. Flesch Reading Ease scores ranged from 72.52 to 91.36 (mean = 82.75), consistent with texts in the "fairly easy" to "easy" range. Average sentence length ranged from 5.86 to 11.32 words per sentence (mean = 7.85).

Fluency and Coherence. Entity continuity scores ranged from 0.10 *(How About a Wig?)* to 0.81 *(The Business Man)*, with a mean of 0.62 (SD = 0.24). The notably low entity continuity for *How About a Wig?* (0.10) reflects its narrative structure, which alternates between two characters rather than maintaining a single referential chain — a legitimate stylistic choice rather than a coherence failure. Adjacent sentence similarity ranged from 0.16 to 0.26 (mean = 0.20), indicating modest local topical continuity across all stories. Sentence length standard deviation ranged from 1.70 to 4.51, with longer stories *(The Post Office, Jim and Antonia)* showing greater structural variability.

Variation by Source Book. My Antonia stories had the highest mean FK Grade Level (4.17) and longest mean sentences (9.27 words), driven primarily by The Post Office and Jim and Antonia, which are the two longest and most complex stories in the corpus. Stories derived from A White Heron had the lowest mean FK Grade Level (3.18) and shortest sentences (6.39 words), as well as the lowest entity continuity (0.57), reflecting the shorter, more episodic narrative structure of that book's stories. The Great Gatsby stories fell between these two (FKGL = 3.65, 7.90 words/sentence). These between-book differences in the human baseline

Table 3: Human-Written Story Baseline Metrics

Story Title	Simplicity				Fluency & Coherence		
	FKGL	FRE	WPS	SPW	Ent. Cont.	Adj. Sim.	SL SD
Jim and Antonia	5.18	76.25	10.29	1.420	0.778	0.212	3.91
Mary and Martha	4.76	72.52	6.50	1.510	0.800	0.225	1.84
The Post Office	4.05	86.16	11.32	1.291	0.667	0.159	4.51
The Garden	3.27	82.63	6.19	1.394	0.600	0.259	1.94
Nick Changes His Mind	3.53	80.54	6.04	1.420	0.600	0.171	1.95
The Business Man	1.97	91.36	5.86	1.295	0.810	0.159	2.42
The Valley of Ashes	3.76	85.44	9.75	1.318	0.632	0.279	4.44
How About a Wig?	2.80	87.11	6.82	1.333	0.095	0.165	1.70
Mean	*3.67*	*82.75*	*7.85*	*1.373*	*0.623*	*0.204*	*2.84*
SD	*1.03*	*6.14*	*2.22*	*0.077*	*0.230*	*0.047*	*1.23*

FKGL = Flesch-Kincaid Grade Level; FRE = Flesch Reading Ease; WPS = Average Words per Sentence; SPW = Average Syllables per Word; Ent. Cont. = Entity Continuity; Adj. Sim. = Adjacent Sentence Similarity; SL SD = Sentence Length Standard Deviation.

are important context for interpreting the cross-validation results further.

4.2. LLM Generated Story Characteristics

Table 5 presents the descriptive statistics for the LLM Generated stories across two evaluation criteria, aggregated by target book. Table 6 presents the cross-validation results, reporting the percentage of generated stories in each fold whose metric values fell within the min-max range of the 7 test stories.

Simplicity. Generated stories were on average simpler than the human-written baseline. Mean FK Grade Level was 2.94 (SD = 1.19) compared to 3.67 (SD = 1.03) for human stories. Average words per sentence was 5.74 for generated stories as compared to 7.85 for human stories. Generated stories were also substantially more uniform in sentence length (SD = 1.25) than human stories (SD = 2.84), reflecting close adherence to the sentence length constraint in the prompt. By target book, *A White Heron* stories were simplest (FKGL = 2.71) and *My Antonia* stories most complex (FKGL = 3.24), mirroring the pattern in the human baseline (See Table 5).

Cross-validation results (Table 6) showed that simplicity metrics had the highest overlap with the test set range for each fold. For FK Grade Level, 67.5% of generated stories fell within the test set range (SD = 17.6 across folds), and similarly 67.5% for average syllables per word (SD = 10.9). Flesch Reading Ease showed 65.0% overlap (SD = 9.3). Average words per sentence had relatively lower overlap at 40.0%, with a mean delta of -2.10 words below the test set mean, confirming that generated sentences were consistently shorter than the simplest human stories in most folds. This metric also showed the highest variability across folds (SD = 32.1), ranging from 6.7% (*The Business Man*) to 93.3% (*The Post Office*), suggesting that exemplar sentence length strongly influenced generation output. Fold-level FKGL means ranged from 2.24 (*The Garden* as exemplar story) to 3.59 (*The Valley of Ashes* as exemplar story), indicating additional sensitivity to training exemplar complexity.

Fluency and Coherence. Generated stories showed lower coherence than human-written stories across all three measures. Entity continuity was 0.39 (SD = 0.29) for generated stories versus 0.62 (SD = 0.23) for human stories, and adjacent sentence similarity showed the largest gap (generated: 0.093; human: 0.204).

The cross-validation results quantify this gap more precisely. Adjacent sentence similarity had the lowest overlap with the test set range of any metric: only 5.0% of generated stories fell within the human range (SD = 7.8), with 0.0% overlap in five of eight folds (Table 6). Sentence length standard deviation showed similarly low overlap at 12.5% (SD = 15.7). Entity continuity showed moderate overlap at 64.2%, but with high variability across folds (SD = 28.7): overlap reached 93.3% when *Jim and Antonia* was the exemplar but dropped to 0.0% when *How About a Wig?* was the exemplar. This variability is driven by the unusual entity continuity profile of *How About a Wig?* (0.095), which alternates between two characters rather than maintaining a single referential chain; when this story is excluded from the test set (Fold 7), the test set range narrows substantially, causing all generated stories to fall outside it. The consistently near-zero overlap for adjacent sentence similarity across folds indicates that the coherence gap is a systematic limitation of the LLM's generation behavior rather than an artifact of particular training exemplars.

Thematic Adherence. Embedding-based se-

Table 4: Human-Written Story Baseline Metrics by Source Book

	Simplicity				Fluency & Coherence		
Source Book	FKGL	FRE	WPS	SPW	Ent. Cont.	Adj. Sim.	SL SD
A White Heron	3.18	83.66	6.39	1.379	0.568	0.183	1.99
My Antonia	4.17	81.68	9.27	1.368	0.682	0.210	3.45
The Great Gatsby	3.65	82.99	7.90	1.369	0.616	0.225	3.20

Stories per book: *A White Heron* (n = 3), *My Antonia* (n = 3), *The Great Gatsby* (n = 2).

Table 5: LLM-Generated Story Metrics by Training Exemplar and Target Book

	Simplicity				Fluency & Coherence			Theme
Target Book	**FKGL**	**FRE**	**WPS**	**SPW**	**Ent. Cont.**	**Adj. Sim.**	**SL SD**	**Sem. Sim.**
A White Heron	2.71	86.53	6.11	1.349	0.333	0.110	1.275	0.370
My Antonia	3.24	82.08	5.77	1.406	0.322	0.079	1.251	0.422
The Great Gatsby	2.86	84.11	5.35	1.387	0.529	0.089	1.224	0.385
Mean	*2.94*	*84.24*	*5.74*	*1.380*	*0.395*	*0.093*	*1.250*	*0.392*
SD	*1.19*	*8.89*	*0.95*	*0.109*	*0.285*	*0.039*	*0.376*	*0.091*

Sem. Sim. = Theme-Text Semantic Similarity; Each cell is contains mean values computed across all 40 stories (8 folds $\times$ 5 stories) generated for each target book.

mantic similarity between each story's declared theme statement and its text was moderate, with a mean of 0.39 (SD = 0.09, range: 0.20-0.64). This indicates that generated stories generally reflected their declared themes at the semantic level. Similarity was comparable across target books: A White Heron (mean = 0.37), The Great Gatsby (mean = 0.39), and My Antonia (mean = 0.42).

Within-book theme diversity showed that the five stories generated for the same book within a fold were moderately distinct from one another, with mean pairwise TF-IDF similarity of 0.44 (SD = 0.13, range: 0.25-0.82) across all book-fold combinations. This indicates the LLM explored different thematic territory across the five stories rather than collapsing to a single dominant theme, though some theme overlap was present. The maximum pairwise similarity between any two theme statements within a book-fold combination reached 1.00 in two instances (both involving *My Antonia* as the target book), indicating occasional identical theme statements.

Within-fold theme diversity showed that the 15 stories generated across all three target books within a fold were distinct from one another (mean pairwise TF-IDF = 0.40, SD = 0.07), indicating that theme generation was driven more by story-level variation than by target book. Maximum pairwise similarity within a fold reached 1.00 in four of eight folds, again reflecting occasional identical theme statements, but mean similarity remained consistently low across folds (range: 0.33-0.53), suggesting no systematic cross-book theme bleed.

Same-Source vs. Cross-Source Prompting. Same-source prompting (i.e., the source book of the exemplar matching the target book) produced stories that more frequently fell within the test set range across most metrics. The largest differences were observed for entity continuity (82.5% of same-source stories vs. 55.0% of cross-source stories in the test set range), average words per sentence (60.0% vs. 30.0%), and adjacent sentence similarity (12.5% vs. 1.2%). Same-source prompting also consistently produced lower FK Grade Level than cross-source prompting for all three target books.

The effect on coherence was inconsistent across books. For *A White Heron*, same-source prompting produced higher entity continuity than cross source (0.416 vs. 0.283), suggesting that a same-book exemplar better scaffolded referential coherence. For *The Great Gatsby*, the pattern reversed: cross-source prompting produced higher entity continuity (0.566 vs. 0.418). Theme semantic similarity was comparable across both conditions for all three books (differences $\leq$ 0.06), indicating that exemplar source had no meaningful effect on thematic coherence.

5. Discussion

The simplicity results confirm that one-shot prompting with explicit sentence length and vocabulary constraints can reliably produce accessible text, consistent with prior findings that LLMs respond well to structural specifications in educational content generation (Laverghetta Jr. and Licato, 2023). Cross-validation showed that approximately two-thirds of generated stories fell within the test set range for simplicity metrics, and that fold-level FKGL means varied depending on the training exemplar. The sensitivity to exemplar complexity sug-

Table 6: Cross-Validation: Percentage of Generated Stories Within Test set Range, by Fold

		Simplicity				Fluency & Coherence		
Fold	**Training Exemplar Used**	**FKGL**	**FRE**	**WPS**	**SPW**	**Ent. Cont.**	**Adj. Sim.**	**SL SD**
0	Jim and Antonia	73.3	66.7	80.0	66.7	93.3	0.0	13.3
1	Mary and Martha	60.0	60.0	20.0	46.7	80.0	6.7	0.0
2	The Post Office	73.3	60.0	93.3	60.0	86.7	0.0	46.7
3	The Garden	53.3	66.7	26.7	80.0	60.0	0.0	0.0
4	Nick Changes His Mind	80.0	73.3	13.3	66.7	66.7	0.0	0.0
5	The Business Man	33.3	46.7	6.7	80.0	60.0	13.3	13.3
6	The Valley of Ashes	86.7	73.3	26.7	66.7	66.7	20.0	6.7
7	How About a Wig?	80.0	73.3	53.3	73.3	0.0	0.0	20.0
Mean		67.5	65.0	40.0	67.5	64.2	5.0	12.5
SD		17.6	9.3	32.1	10.9	28.7	7.8	15.7

Each cell reports the percentage of 15 generated stories in that fold whose metric value falls within the min–max range of the 7 test set stories. Mean and SD are computed across the 8 fold-level percentages.

gests that strategic exemplar selection could serve as a practical lever for targeting specific readability bands in operational use. However, the uniformity of generated sentence lengths (SD = 1.25 vs. 2.84 for human stories) suggests that the model might have interpreted the prompt's sentence length constraint as a target rather than a ceiling, producing text that is structurally monotonous. Whereas, human-written DLM stories varied sentence length deliberately to maintain reader engagement and signal narrative transitions—properties that matter for students who rely on predictable but not rigid textual patterns (Erickson and Geist, 2016). Future prompt designs should distinguish between a maximum sentence length and the expectation of natural variation within that bound.

The drop in coherence for LLM-generated stories in comparison to human-authored stories is the most consequential finding: only 5.0% of generated stories achieved adjacent sentence similarity within the test set range, and this near-zero overlap held across folds (0.0% in five of eight). This does not mean that 95% of generated stories are entirely incoherent; rather, it means that the degree to which consecutive sentences share overlapping vocabulary and content was consistently lower than what human experts produced, suggesting that LLM-generated stories tend to introduce new referents between sentences more than human-written DLM stories. However, students with SCD referential continuity is not a stylistic preference but a comprehension necessity. These readers depend on explicit cues such as repeated character names, pronoun chains, and topical bridges between sentences to track who is doing what across a story. When those cues are absent, even individually simple sentences become difficult to integrate into a coherent mental model of the narrative. Incorporating explicit coherence instructions into the generation prompt, following cognitively motivated generation strategies (Chamovitz and Abend, 2022)(e.g., requiring that each sentence share at least one referent with its predecessor, or that character names be reused across non-adjacent sentences) could address this gap without sacrificing simplicity.

Thematic adherence was moderate, with generated stories generally reflecting their declared themes at the semantic level and maintaining adequate within-book diversity. However, the occasional production of identical theme statements within a fold (4 of 8 folds) suggests that the model's theme generation draws from a constrained latent space of "accessible story themes" rather than engaging deeply with the source material's thematic range. In practice, this means that while individual stories may appear thematically appropriate, a set of five stories for the same book may lack the substantive differentiation needed to support distinct assessment items targeting different reading comprehension skills. Human review at the theme-selection stage, or prompt modifications that provide explicit thematic anchors drawn from different chapters or subplots of the source book, could mitigate this collapse.

6. Conclusion

In conclusion, our findings suggest that LLMs can serve as a useful drafting tool within the DLM development pipeline, but not as a replacement for human authoring. Generated stories would need revision by accessibility specialists to address coherence before entering the review stages that are fundamental to DLM quality assurance (DLM Consortium, 2024). This aligns with the broader consensus that human-in-the-loop oversight is essential for AI-generated educational content in high-stakes contexts (Clark et al., 2025; Tan et al., 2025). Future work should target the coherence gap through prompt engineering (e.g., explicit instructions for entity reuse and topical connectivity), incorporate expert human review alongside automated metrics,

expand the pipeline to additional grade levels and text types, and evaluate whether students and educators perceive meaningful differences between human-written and LLM-generated stories in assessment contexts.

7. Ethical Considerations and Limitations

Several limitations warrant careful consideration. First, the human baseline comprised only 8 stories compared to 120 LLM-generated stories, creating an asymmetry that limits the precision of comparison. Second, the evaluation framework used in this study has not been reviewed by DLM operational staff, nor have the AI-generated stories themselves been evaluated through DLM's internal review processes. Two criteria central to DLM's own quality standards are absent from our framework: (a) every DLM text must contain sufficient testable points aligned to a specific node to support the development of approximately five distinct items measuring the same skill (because we primarily focused on assessing LLMs' story generation capabilities), and (b) Bias, sensitivity and people-first language flags (which we plan to address in future work that extends the evaluation framework). It is important to note that even though we propose the use of readability formulas such as Flesch-Kincaid, DLM relies more heavily on expert human judgment which capture nuances missed by automated metrics. Consequently, the differences observed in our automated metrics should not be interpreted as evidence that LLM-generated stories surpass human-written DLM texts in quality. Rather, the differential may reflect that the human-written texts met a categorically different and more comprehensive standard of quality than what our automated metrics capture. Third, we only experiment with one model (OpenAI's GPT-4o) for story generation and assumed that the LLM had prior knowledge of the canonical source books used in this study, an assumption that may not hold across models or when extending this work to less widely known, multilingual, or culturally specific literary sources. Fourth, the generation setup is intentionally minimal — one-shot prompting with a single human-written exemplar due to limited availability of human written examples which limits the strength of conclusions that can be drawn about LLM capability for this task more broadly. Future work can compare our results with alternative prompting strategies such as few-shot prompting, chain-of-thought, constrained decoding, or structured prompting.

Given these limitations, a critical next step is to have the generated stories evaluated by human experts using both the automated metrics proposed here and DLM's own review criteria, including testability, accessibility, and bias and sensitivity standards. Such a validation study would allow assessment of the reliability and validity of the proposed evaluation framework, determine whether the automated metrics align with expert judgment, and establish whether LLM-generated stories can meet the comprehensive quality standards required for operational use. More broadly, ethical deployment of AI-generated content in high-stakes assessment contexts demands that generated texts undergo the same rigorous multi-stage review process applied to human-authored stories, ensuring that no content reaches students without thorough human oversight.

8. Lay Summary

Reading comprehension tests for students with significant cognitive disabilities(SCD) require specially written short stories that use simple words, short sentences, and clear, easy-to-follow narratives. Writing these stories takes a great deal of time and expertise. This study explored whether large language models could help by generating first drafts of these stories. We gave (GPT-4o) a single example of a human-written story and asked it to produce new short stories inspired by well-known books like "The Great Gatsby" and "My Antonia", but written in a simple and accessible way. We then measured how well the AI-generated stories compared to stories written by human experts across three criteria: (1) how easy they were to read, (2) how well each sentence connected to the next, and (3) whether the stories stayed focused on a single theme. We found that the AI was good at keeping the language simple, but consistently struggled to write sentences that flowed naturally from one to the next in the way that human experts did — a feature that is especially important for students with SCD who rely on clear, predictable language to understand what they read. Stories generally stayed on topic, though the AI sometimes repeated the same themes across different stories. Overall, our findings suggest that AI can be a useful starting point in the story-writing process, but that human expert review and revision remain essential before these stories could be used in real assessments.

9. Acknowledgments

This research was made possible through the generous Research Fellowship support of Accessible Teaching, Learning, and Assessment Systems (ATLAS) at the University of Kansas.

10. References

Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. 2020. Asset: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4668–4679. Association for Computational Linguistics.

Catarina G. Belem, Parker Glenn, Alfy Samuel, Anoop Kumar, and Daben Liu. 2025. Readability reconsidered: A cross-dataset analysis of reference-free metrics. *arXiv preprint arXiv:2510.15345*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Elad Chamovitz and Omri Abend. 2022. Cognitive simplification operations improve text simplification. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 241–265.

A. K. Clark, A. Hirt, D. Whitcomb, W. J. Thompson, M. Wine, and M. Karvonen. 2025. Artificial intelligence in science and mathematics assessment for students with disabilities: Opportunities and challenges. *Education Sciences*, 15(2):233.

DLM Consortium. 2024. 2023–2024 technical manual—dlm alternate assessment system. Technical report, University of Kansas, Accessible Teaching, Learning, and Assessment Systems (ATLAS).

Karen Erickson and Lori A. Geist. 2016. The profiles of students with significant cognitive disabilities and complex communication needs. *Augmentative and Alternative Communication*, 32(3):187–197.

Yi Feng, Mingyang Song, Jiaqi Wang, Zhuang Chen, Guanqun Bi, Minlie Huang, Liping Jing, and Jian Yu. 2025. Ss-gen: a social story generation framework with large language models. AAAI'25/IAAI'25/EAAI'25. AAAI Press.

Meagan Karvonen and A. K. Clark. 2019. Students with the most significant cognitive disabilities who are also english learners. *Research and Practice for Persons with Severe Disabilities*, 44(2):71–86.

Meagan Karvonen, A. K. Clark, C. Carlson, S. Wells Moreaux, and J. Burnes. 2021. Approaches to identification and instruction for students with significant cognitive disabilities who are english learners. *Research and Practice for Persons with Severe Disabilities*, 46(4):223–239.

J. Peter Kincaid, Robert P. Fishburne, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command, Millington, TN, Research Branch.

Anthony Laverghetta Jr. and John Licato. 2023. Generating better items for cognitive assessments using large language models. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 414–428. Association for Computational Linguistics.

Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. 2022. MUSS: Multilingual unsupervised sentence simplification by mining paraphrases. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1651–1664, Marseille, France. European Language Resources Association.

Brooke Nash, A. K. Clark, and Meagan Karvonen. 2016. First contact: A census of students taking the dynamic learning maps alternate assessment. Technical report, University of Kansas, Center for Educational Testing and Evaluation.

OpenAI. 2023. Gpt-4 technical report.

G. Raffloer and Melanie C. Green. 2025. Of love & lasers: Perceptions of narratives by ai versus human authors. *Computers in Human Behavior: Artificial Humans*, 5:100168.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992. Association for Computational Linguistics.

B. Tan, N. Armoush, E. Mazzullo, O. Bulut, and M. Gierl. 2025. A review of automatic item generation techniques leveraging large language models. *International Journal of Assessment Tools in Education*, 12(2):317–340.

Martha L. Thurlow, S. S. Lazarus, E. D. Larson, D. A. Albus, K. K. Liu, and E. Kwong. 2017.

Alternate assessments for students with significant cognitive disabilities: Participation guidelines and definitions. Technical report, University of Minnesota, National Center on Educational Outcomes.

Martha L. Thurlow, Y. Wu, Rachel F. Quenemoen, and E. Towles. 2016. Characteristics of students with significant cognitive disabilities. Technical report, National Center and State Collaborative.

Laura Vásquez-Rodríguez, Matthew Shardlow, Piotr Przybyła, and Sophia Ananiadou. 2023. Document-level text simplification with coherence evaluation. In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 85–101, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Shira Yalon-Chamovitz. 2009. Invisible access needs of people with intellectual disabilities: A conceptual model of practice. *Intellectual and Developmental Disabilities*, 47(5):395–400.

A. Story Generation Prompt

The following system and user prompts were used to generate stories via the OpenAI API (GPT-4o, temperature = 0.4). The `{target_book}` and `{n}` placeholders are filled at runtime per fold.

System Prompt

You are an English Language Arts story writer producing accessible short stories for Grade 11–12 students with cognitive disabilities.

Rules:

- Produce short stories in plain, simple sentences inspired by the book `{target_book}`, but adapted for students.
- Each story should be between 15–30 sentences long.
- Use simple vocabulary.
- Characters and settings should relate to themes from `{target_book}`.
- For each story, provide a theme statement in the form: *"This story explores [theme]."*
- The theme statement must be reflected consistently throughout the story text.
- Output MUST be a single JSON array of exactly `{n}` objects with keys: `story_id`, `book_inspiration`, `grade`, `story_text`, `title`, `theme_statement`.
- Do NOT include any prose, explanations, or code fences. JSON array only.

User Prompt

`EXEMPLAR={exemplar_json}`

Now generate `{n}` independent stories inspired by "`{target_book}`" following the exemplar structure and style constraints. Return ONLY a JSON array with `{n}` story objects. No wrapper. No extra text.

B. Example Stories

B.1. Human-Written Story

Source Book: *My Antonia*
Story Title: Jim and Antonia
Grade: 11–12

1. Jim was a boy.
2. Jim was ten years old.
3. Jim moved to Nebraska to stay with the Burdens.
4. The Burdens were Jim's grandparents.
5. The Shimerdas were the Burdens' new neighbors.
6. The Shimerdas did not have a garden yet.
7. Jim's grandmother decided to pack food into an old wagon and take it to the Shimerdas.
8. Jim and his grandmother packed potatoes and pork to take to the Shimerdas.
9. Jim and his grandmother packed bread, butter, and pumpkin pie to take to the Shimerdas.
10. Jim and his grandmother rode in the wagon to the Shimerdas' house.
11. Jim and his grandmother arrived at the Shimerdas' house.
12. It was a small house built into the side of a hill.
13. Jim and his grandmother met Mr. and Mrs. Shimerda and their daughter Antonia.
14. The Shimerdas did not speak very much English.
15. Jim needed someone to tell him what the Shimerdas said.
16. Jim and Antonia played together in a field while Jim's grandmother stayed with Mr. and Mrs. Shimerda.
17. Antonia wanted to learn the English words for things that she saw.
18. Antonia pointed to the things she wanted to learn how to say.
19. Jim told Antonia the English words for the things she pointed to.
20. Antonia pointed to the sky, a tree, and Jim's eyes.
21. Jim told her the English words for all the things she pointed to.
22. Antonia learned many new English words while she played with Jim.
23. When it was time for Jim and his grandma to leave, Mr. Shimerda asked Jim's grandmother for a favor.
24. Mr. Shimerda asked Jim's grandmother to teach Antonia how to speak English.
25. Jim's grandmother said she would.
26. Antonia was excited to learn more English words.
27. Jim was excited to have a new friend.

B.2. LLM-Generated Story

Example 1

Target Book: *A White Heron*
Story Title: Sylvia and the Heron
Grade: 11–12
Theme Statement: This story explores the importance of protecting nature.
Training Exemplar: Jim and Antonia

1. Sylvia lived with her grandmother in the woods.
2. Sylvia loved the woods and the animals that lived there.
3. One day, Sylvia met a man in the woods.
4. The man was a hunter. He was looking for a white heron.
5. The man asked Sylvia if she had seen the heron.
6. Sylvia knew where the heron lived.
7. The man offered Sylvia money to help him find the heron.
8. Sylvia liked the man, but she loved the heron more.
9. Sylvia climbed a tall tree to see the heron's nest.
10. Sylvia saw the heron flying in the sky.
11. Sylvia felt happy watching the heron.
12. Sylvia decided not to tell the man where the heron was.
13. Sylvia wanted to protect the heron.
14. The man left the woods without finding the heron.
15. Sylvia was glad she kept the heron safe.

Example 2

Target Book: *My Antonia*
Story Title: The New Home
Grade: 11–12
Theme Statement: This story explores family togetherness.
Training Exemplar: Jim and Antonia

1. Anna and her family moved to a new place.
2. Their new home was on a big farm.
3. Anna's father worked hard to build their house.
4. Anna's mother planted a garden.
5. Anna helped her mother plant seeds in the soil.
6. Anna liked to watch the plants grow.
7. Anna's little brother played in the fields.
8. Anna and her brother found many bugs and flowers.
9. Anna's family did not have much money.
10. Anna's mother made clothes from old fabric.
11. Anna learned how to sew from her mother.
12. Anna made a dress for her little brother.
13. Anna's father taught her how to care for the animals.
14. Anna fed the chickens and cows every day.
15. Anna loved her new home on the farm.
16. Anna and her family worked hard together.
17. Anna felt happy and safe with her family.
18. Anna knew they would be okay in their new home.

Example 3

Target Book: *The Great Gatsby*
Story Title: Nick and the Party
Grade: 11–12
Theme Statement: This story explores social connections.
Training Exemplar: Jim and Antonia

1. Nick was a quiet man.
2. Nick lived next to Jay.
3. Jay invited Nick to a party.
4. Nick was excited to go.
5. The party had bright lights.
6. There was music and dancing.
7. Nick met many people.
8. Nick talked to a man named Tom.
9. Tom was loud and funny.
10. Nick liked talking to Tom.
11. Nick also met Daisy.
12. Daisy was kind and friendly.
13. Nick felt happy at the party.
14. Nick made new friends.
15. Nick hoped to go to more parties.

Evaluating Transformer Model Family Representations Through Automated Essay Scoring

Akchay Ozten, Rodrigo Wilkens
University of Exeter
{ao478, r.wilkens}@exeter.ac.uk

Abstract

Large Language Models have become central to Automated Essay Scoring (AES), typically through fine-tuned transformer encoders or prompt-based applications of decoder models. However, the representational capacity of decoder models as frozen embedding extractors remains largely unexplored. In this paper, we present a controlled comparison between encoder and decoder transformer embeddings for prompt-agnostic AES. Using regression models, we evaluate frozen representations across two English datasets. We analyzed scaling effects and the impact of integrating explicit linguistic features in hybrid configurations. Our results show that decoder embeddings consistently outperform encoder embeddings in embedding-only settings, with gains generalizing across holistic essay scoring and proficiency prediction. Scaling effects are modest, and hybrid models that combine contextual embeddings with linguistic features yield further improvements. Notably, frozen decoder embeddings achieve performance competitive with a fine-tuned BERT. These findings highlight the importance of representation-level properties in essay scoring.

Keywords: Automated Essay Scoring, Autoregressive Models, Transformer Representations, CEFR Classification, Hybrid NLP Models

1. Introduction

Automated Essay Scoring (AES) is the use of machine learning (ML) and natural language processing (NLP) techniques to evaluate and grade written essays automatically (Taghipour and Ng, 2016; Shermis and Burstein, 2013). AES systems provide consistent, objective, and scalable assessments, reducing the workload of human graders while potentially offering rapid feedback to students (Klebanov and Madnani, 2022). Essays are typically written in response to specific prompts, and scoring requires assessing multiple dimensions of writing quality, including coherence, fluency and grammatical accuracy.

Early AES systems relied on manually engineered features such as lexical diversity, syntactic complexity, and surface-level readability metrics (Page, 1966; Attali and Burstein, 2006; Foltz et al., 1999; Zesch et al., 2015). While these approaches achieved moderate correlations with human raters, they struggled to capture deeper semantic and discourse-level properties and were vulnerable to superficial manipulation (Perelman, 2014; Shermis and Burstein, 2013). Subsequent neural approaches based on CNNs and RNNs reduced reliance on feature engineering and improved representation learning (Taghipour and Ng, 2016; Wang et al., 2018), yet they remained limited in modeling longer essays.

The introduction of the transformer architecture transformed Natural Language Processing (NLP) by enabling efficient modeling of long-range dependencies through self-attention (Vaswani et al., 2017). Encoder-based models have since become dominant in AES research. Fine-tuning pre-trained encoders on scoring datasets has yielded strong performance (Rodriguez et al., 2019; Mayfield and Black, 2020), and hybrid approaches combining contextual embeddings with hand-crafted features have further improved results (Dasgupta et al., 2018; Uto et al., 2020). Such hybrid architectures leverage both deep contextual representations and explicit linguistic signals.

Decoder-based generative models have gained attention in AES through prompt-based evaluation strategies (Mizumoto and Eguchi, 2023; Yancey et al., 2023; Stahl et al., 2024). These models generate scores conditioned on rubrics or other prompted information. While prompt-based methods have shown promising performance, they are limited to generative behavior.

Despite the growing influence of generative models, their internal embeddings have been largely overlooked in NLP and AES research. In particular, there has been limited investigation into whether frozen decoder representations, used independently of prompting, encode signals relevant to writing quality. This study investigates whether embeddings extracted from decoder models, when used as fixed representations or in a hybrid framework, can enhance the performance of AES systems. To evaluate robustness, we conduct a prompt-agnostic comparison across two assessment settings. We assess whether any observed representational advantages generalize across related evaluation tasks. The study is guided by the following research questions:

RQ1 Do frozen decoder embeddings improve per-

Proceedings of the Joint Workshop on Readability and Text Simplification (READIxTSAR) @ LREC 2026, pages 142–150
11 May 2026.

formance over encoder embeddings in prompt-agnostic AES?

RQ2 How does model size influence the effectiveness of decoder representations for scoring?

RQ3 Do linguistic features remain complementary when combined with decoder embeddings?

The main contribution of this study is a systematic comparison of encoder and decoder embeddings as frozen representations for AES and related assessment tasks. Our results show that decoder embeddings consistently outperform encoder embeddings in this setting and that these gains generalize across datasets. Furthermore, we demonstrate that hand-crafted linguistic features provide complementary information when integrated with transformer-based representations.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 details the methodology. Section 4 presents the results, followed by discussion in Section 5 and concluding remarks in Section 6.

2. Related Work

2.1. Transformer-Based Approaches

Transformer architectures have become central to Automated Essay Scoring due to their ability to model long-range dependencies and contextual interactions (Vaswani et al., 2017). Encoder-based models, such as BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), XLNet (Yang et al., 2019), and DeBERTa (He et al., 2020, 2021), have been widely adopted in AES (Devlin et al., 2018; Liu et al., 2019; He et al., 2020, 2021). Fine-tuning pre-trained encoders for essay scoring has consistently demonstrated improvements over earlier neural approaches (Rodriguez et al., 2019; Mayfield and Black, 2020). In these systems, scoring is formulated as a supervised regression or classification task, with a task-specific head trained on top of contextualized document representations.

A second line of work integrates encoder representations within hybrid architectures that concatenate contextual embeddings with linguistic features (Dasgupta et al., 2018; Uto et al., 2020). These features typically capture lexical richness, syntactic complexity or discourse-level statistics. Empirical results suggest that hybrid systems often outperform purely neural models. However, both fine-tuning and hybrid approaches have predominantly relied on encoder-derived document representations, typically extracted via pooled outputs (e.g., the *[CLS]* token).

2.2. Prompt-Based and Decoder Architecture

From a different perspective, large autoregressive decoder models, such as GPT (Radford et al., 2018), Mistral (Jiang et al., 2023), Qwen (Bai et al., 2023), and DeepSeek (Bi et al., 2024), have achieved strong performance in generative language tasks. Their flexibility through prompting has led to growing interest in applying them to assessment settings (Liu et al., 2023). Mizumoto and Eguchi (2023) evaluated GPT-3 for rubric-based essay scoring, Yancey et al. (2023) examined GPT-3.5 and GPT-4, and Stahl et al. (2024) investigated prompting strategies using Mistral 7B. These works report performance competitive with traditional supervised AES systems on datasets such as TOEFL11, demonstrating the potential of generative inference for automated scoring.

Methodologically, prompt-based AES differs fundamentally from encoder fine-tuning. Rather than training a regression head over fixed document embeddings, decoder models are evaluated in generative inference mode, conditioned on prompts, scoring rubrics, or example essays. Consequently, scoring performance may depend on prompt formulation and decoding strategies (Liu et al., 2023). This makes it difficult to isolate intrinsic representational capacity from inference-time adaptation effects.

Encoder and decoder architectures are also trained under distinct pre-training objectives. Encoder models rely on masked language modeling (MLM) (Devlin et al., 2018), whereas decoder models are optimized using autoregressive next-token prediction (Radford et al., 2018; Brown et al., 2020). These objectives impose different structural constraints on learned representations and have been shown to induce distinct inductive biases and internal geometries (Rogers et al., 2020; Belinkov and Glass, 2019).

Despite the increasing use of decoder models in prompt-based AES, prior research has largely examined them in generative inference settings rather than as sources of document-level embeddings for supervised scoring. To our knowledge, the use of decoder-derived representations as fixed embedding features in AES remains largely unexplored.

3. Methodology

To address the research questions outlined in Section 1, we adopt a controlled comparative framework in which the only systematically varied component is the source of transformer-based representations. Specifically, we compare embeddings extracted from encoder-based and decoder-based transformer models under identical downstream

conditions.

3.1. Corpora

This study employs two English datasets representing related but distinct assessment settings. Corpus statistics are summarized in Table 1.

The first dataset is the AES2 benchmark (Crossley et al.), released as part of the Learning Agency Lab Automated Essay Scoring 2.0 competition. It contains 17,307 essays scored on a 1-6 ordinal scale. Unlike the earlier ASAP dataset[1], AES2 does not provide explicit prompt labels, encouraging prompt-agnostic evaluation and broader generalization.

The second corpus is the Common European Framework of Reference for Languages (CEFR) European Language Grid (ELG) dataset in English (Breuker, 2023). It is available as part of the UniversalCEFR project (Imperial et al., 2025) in HuggingFace[2]. It contains 712 essays labeled according to CEFR proficiency levels. To mitigate class imbalance, adjacent sublevels (e.g., A1-/A1+, A2-/A2+) are merged, resulting in six levels aligned with the AES scoring scale. While AES2 evaluates holistic essay quality, the ELG dataset captures the result of a descriptor-based method.

Level	#Essays	#Tokens
A1	1,252	274 (110)
A2	4,723	265 (97)
B1	6,280	361 (102)
B2	3,926	483 (109)
C1	970	636 (135)
C2	156	778 (162)

(a) AES2

Level	#Essays	#Tokens
A1	25	255 (14)
A2	141	216 (71)
B1	206	282 (67)
B2	174	423 (181)
C1	97	640 (324)
C2	69	775 (239)

(b) ELG

Table 1: Distribution of essays and average token counts (mean with standard deviation in parentheses) across CEFR levels

[1] https://www.kaggle.com/c/asap-aes

[2] https://huggingface.co/datasets/UniversalCEFR/elg_cefr_en

3.2. Representation Extraction

We evaluate multiple encoder and decoder transformer models obtained from the HuggingFace library (Wolf et al., 2019). Encoder models include BERT and its variants (Devlin et al., 2018; He et al., 2020, 2021), while decoder models include Mistral (Jiang et al., 2023), Qwen (Bai et al., 2023), LLaMA (Llama Team, AI @ Meta, 2024), and DeepSeek (Bi et al., 2024). These models vary in parameter size, enabling analysis of scaling effects.

Embeddings are derived from the final hidden layer of each transformer model. For encoder models, we extract the embedding corresponding to the *[CLS]* token, which is conventionally used as a global sequence representation in classification and regression tasks (Devlin et al., 2018).[3] This representation serves as a compact essay-level embedding.

Decoder models do not employ a dedicated classification token. To obtain a fixed-size representation, we apply mean pooling over the hidden states of all tokens in the final layer. This strategy yields a length-invariant summary while preserving contextual information learned through autoregressive training. Although this approach produces a representation similar to that obtained for the decoders, one might argue that there is an asymmetry in the comparisons. To address this perspective, we also evaluated the BERT and DeBERTa models using mean pooling.

This study employed 15 transformer models, 5 encoders and 10 decoders, from Hugging Face to extract embeddings. These are small to medium models. Table 2 provides the details including the size of the embedding vectors.

All transformer models remain frozen throughout the experiments. No task-specific fine-tuning is performed in the embedding-based configurations.

3.3. Linguistic Features

To evaluate complementarity between contextual embeddings and explicit linguistic information (RQ3), we extract a set of features using the *TextDescriptives* library (Hansen et al., 2023). These features include descriptive statistics, lexical richness indicators, syntactic complexity measures, dependency distance metrics, and coherence-related indicators. Such features have been shown to correlate with writing quality and proficiency (Zesch et al., 2015; Shermis and Burstein, 2013). When combined with encoder/decoder embeddings, these features form hybrid representations that integrate

[3] The transformers library was used to extract embeddings from both encoder and decoder models. These embeddings are derived from the last hidden state of the model outputs and capture contextualized token representations.

Model	Hugging Face model	Transformer type	Vector size
BERT	google-bert/bert-base-uncased	Encoder	768
BERT Large	google-bert/bert-large-uncased	Encoder	1,024
ModernBERT	answerdotai/ModernBERT-base	Encoder	768
DeBERTa	microsoft/deberta-v3-base	Encoder	768
DeBERTa Large	microsoft/deberta-v3-large	Encoder	1,024
Mistral 7B	mistralai/Mistral-7B-Instruct-v0.3	Decoder	4,096
QWEN3 0.6B	Qwen/Qwen3-0.6B-Base	Decoder	1,024
QWEN3 1.7B	Qwen/Qwen3-1.7B-Base	Decoder	2,048
QWEN3 4B	Qwen/Qwen3-4B-Base	Decoder	2,560
QWEN3 8B	Qwen/Qwen3-8B-Base	Decoder	4,096
QWEN3 14B	Qwen/Qwen3-14B-Base	Decoder	5,120
Llama 1B	meta-llama/Llama-3.2-1B	Decoder	2,048
Llama 3B	meta-llama/Llama-3.2-3B	Decoder	3,072
Deepseek 1.5B	deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B	Decoder	1536
Deepseek 7B	deepseek-ai/DeepSeek-R1-Distill-Qwen-7B	Decoder	3584

Table 2: HuggingFace transformer models used

explicit structural signals with deep contextual encodings.

3.4. Scoring Framework

We adopt a regression-based formulation of essay scoring, enabling the prediction of fine-grained values within the ordinal scale. Treating essay scores as continuous targets avoids rigid class boundaries while allowing ordinal-sensitive evaluation through appropriate metrics.

For all embedding-based experiments, we employ LightGBM and SVM as the downstream regressor.[4] Transformer embeddings, optionally concatenated with linguistic features, are provided as input to the model. LightGBM is a histogram-based gradient boosting framework optimized for high-dimensional feature spaces, making it well suited for transformer embeddings ranging from hundreds to several thousand dimensions. In contrast, a linear SVM serves as a strong linear baseline, effectively assessing the extent to which the frozen embedding space supports linear prediction for the target task.

To address RQ1-RQ3, we evaluate a configuration in which the frozen essay-level embedding extracted from each encoder or decoder model constitutes the sole input to the LightGBM regressor. We extend the embedding-only setup by concatenating the linguistic feature vector to the transformer embedding. The combined representation is then provided as input to the same LightGBM regressor.

[4]For hyperparameter tuning, n_estimators was set to 4,000-6,000 for the AES2 dataset and 1,000-3,000 for the ELG dataset for the LightGBM, and C between 0.01 and 1 for the linear SVM.

3.5. Fine-Tuned Encoder Baseline

In addition to frozen representations, we include a fine-tuned BERT model as a supervised baseline (Rodriguez et al., 2019; Mayfield and Black, 2020). In this configuration, a regression head is added on top of the encoder, and the model is optimized end-to-end for the scoring task (i.e., all layers weights are updated). Training, validation, and evaluation follow the same cross-validation protocol used in the embedding-based experiments. The fine-tuning process was run using the following parameters: learning rate of 0.00002, per device train batch size of 8, per device evaluation batch size of 8, number of train epochs of 5, weight decay of 0.01 and AdamW optimizer.

3.6. Evaluation Protocol

Performance is evaluated using 10-fold cross-validation.[5] The primary metric is Quadratic Weighted Kappa (QWK), which measures agreement between predicted and true scores while penalizing larger discrepancies, and is prioritized as it is widely used in AES benchmarking (Taghipour and Ng, 2016). Accuracy and macro-F1 are also reported, computed after discretizing regression outputs to the nearest valid CEFR score level. Statistical significance between models is assessed using the Wilcoxon signed-rank test across cross-validation folds.

[5]In each fold, we split in 80/10/10.

4. Results

To compare frozen encoder and decoder embeddings under identical downstream conditions, we first evaluate the embedding-only configuration. The results are reported in Table 4.

Across both datasets, decoder-based embeddings consistently outperform encoder-based embeddings. On AES2, all decoder models achieve higher QWK and F1 scores than encoder models under identical regression conditions (Table 4). Statistical testing confirms that the strongest decoder variants significantly outperform the best encoder baselines (Wilcoxon signed-rank test, $p < 0.05$). A similar pattern is observed on the ELG dataset, indicating that the performance advantage of decoder embeddings is not restricted to a single corpus.

First, we compared the performance of the machine learning models (LightGBM vs. SVM), as shown in Tables 4 and 5, respectively. This analysis shows that LightGBM consistently yields better results. We further observe that CLS-based representation consistently outperforms mean pooling across both SVM and LightGBM models (Table 3). This suggests that the way in which document-level representations are constructed has a measurable impact on downstream performance, independent of the regression model. This finding reinforces the importance of representation-level design choices in AES, indicating that not only the model family (encoder vs decoder), but also the aggregation strategy, influences the quality of the resulting embedding space.

Corpus	Model	F1	Accuracy	QWK
AES2	BERT	0.53	0.56	0.69
	DeBERTa	0.58	0.61	0.75
ELG	BERT	0.58	0.59	0.84
	DeBERTa	0.46	0.47	0.76

(a) LightGBM

Corpus	Model	F1	Accuracy	QWK
AES2	BERT	0.55	0.59	0.70
	DeBERTa	0.53	0.57	0.67
ELG	BERT	0.62	0.63	0.85
	DeBERTa	0.52	0.53	0.79

(b) SVM

Table 3: BERT and DeBERTa performance using mean pooling

To evaluate robustness, we examine the consistency of relative model rankings in Table 4. Despite differences in dataset size and scoring criteria, decoder embeddings maintain their advantage across both AES2 and ELG.

We next investigate the relationship between decoder model size and performance (Table 4). While larger models tend to achieve marginally higher scores on AES2, improvements are not consistently statistically significant. On ELG, no monotonic trend is observed. These findings indicate that representational suitability for AES does not scale linearly with parameter count over the evaluated range.

We now compare embedding-only and hybrid configurations. The hybrid results are presented in Table 6. Across all models and both datasets, adding linguistic features yields consistent positive gains. This confirms that contextual embeddings do not fully subsume explicit structural information and that hybrid modeling remains beneficial. Decoder-based hybrid models remain superior to encoder-based hybrids, indicating that representational advantages persist after feature integration.

Finally, we compare frozen embeddings with a fine-tuned BERT (Table 4). Although fine-tuned models achieves competitive results, the strongest frozen decoder embeddings match or exceed its performance. This indicates that representational differences alone can rival supervised adaptation. Moreover, the best fine-tuned model depends on the corpus, likely due to the amount of training data.

In summary, the strongest encoder model (frozen DeBERTa + mean pooling, QWK = 0.74 on AES2; 0.85 on ELG) is consistently outperformed by multiple frozen decoder variants, with Mistral 7B achieving QWK = 0.79 on AES2 and QWEN3/Llama variants reaching up to 0.89 on ELG. The model fine-tuned BERT baseline (DeBERTa QWK = 0.80 on AES2; BERT QWK = 0.85 on ELG) improves over standard encoder embeddings but remains comparable to, and in some cases below, the best frozen decoder models. Overall, the ranking of best-performing approaches is consistent across datasets: frozen decoder embeddings $\geq$ fine-tuned encoder $\geq$ frozen encoder. These results indicate that representation-level differences alone are sufficient to match or exceed supervised adaptation under controlled regression conditions.

5. Discussion

This study evaluated encoder and decoder transformer embeddings for automated essay scoring under strictly controlled downstream conditions. Rather than reiterating performance differences, we discuss their implications for AES research and representation evaluation in NLP.

Recent AES research has converged on fine-tuned encoder architectures as the dominant paradigm (Rodriguez et al., 2019; Mayfield and Black, 2020). Decoder models, when considered, have largely been evaluated through prompt-based scoring frameworks (Mizumoto and Eguchi, 2023; Yancey et al., 2023; Stahl et al., 2024). This division has implicitly positioned encoders as the

Model	AES2			ELG		
	QWK	Accuracy	F1	QWK	Accuracy	F1
BERT	0.70	0.56	0.55	0.82	0.57	0.55
BERT Large	0.71	0.57	0.56	0.76	0.51	0.49
ModernBERT	0.67	0.54	0.53	0.75	0.53	0.51
DeBERTa	0.74	0.59	0.58	0.85	0.62	0.61
DeBERTa Large	0.74	0.59	0.57	0.83	0.60	0.59
Mistral 7B	**0.79**	**0.63**	**0.63**	0.88	0.67	0.66
QWEN3 0.6B	0.75	0.59	0.58	0.87	0.64	0.63
QWEN3 1.7B	0.77	0.60	0.60	0.87	0.65	0.64
QWEN3 4B	0.78	0.61	0.61	0.87	0.65	0.64
QWEN3 8B	0.78	0.61	0.61	0.88	0.67	0.66
QWEN3 14B	0.78	0.61	0.61	**0.89**	0.69	0.68
Llama 1B	0.76	0.59	0.59	0.87	0.65	0.64
Llama 3B	0.77	0.60	0.60	**0.89**	**0.70**	**0.69**
Deepseek 1.5B	0.76	0.60	0.59	0.83	0.59	0.57
Deepseek 7B	0.78	0.62	0.61	0.85	0.63	0.62
BERT-FT	0.77	0.61	0.61	0.85	0.64	0.61
DeBERTa-FT	0.80	0.66	0.65	0.74	0.50	0.43

Table 4: Performance on AES2 and ELG Using LightGBM with the CLS Token and Fine-Tuning

Model	AES2			ELG		
	QWK	Accuracy	F1	QWK	Accuracy	F1
BERT	0.68	0.58	0.54	0.82	0.59	0.58
DeBERTa	0.60	0.51	0.47	0.67	0.44	0.42
Mistral 7B	0.74	0.57	0.54	0.89	0.71	0.69
QWEN3 0.6B	0.73	0.58	0.55	0.87	0.67	0.66
QWEN3 1.7B	0.74	0.60	0.57	0.88	0.67	0.66
QWEN3 4B	0.75	0.60	0.57	0.88	0.68	0.67
QWEN3 8B	0.75	0.60	0.58	0.89	0.69	0.68
QWEN3 14B	0.76	0.60	0.58	0.89	0.71	0.69
Llama 1B	0.74	0.59	0.56	0.88	0.69	0.67
Llama 3B	0.75	0.60	0.57	0.88	0.70	0.69
Deepseek 1.5B	0.75	0.60	0.58	0.87	0.65	0.64
Deepseek 7B	0.76	0.61	0.58	0.87	0.68	0.66

Table 5: Embedding-only performance on AES2 and ELG (SVM)

appropriate representational backbone for scoring, and decoders as generative evaluators.

The present findings complicate that dichotomy. When embeddings are evaluated independently of prompting and fine-tuning, decoder representations consistently match or outperform encoder representations under identical conditions. More broadly, the results indicate that architectural preference in AES has not been systematically stress-tested under controlled representation-level comparisons. The contribution of this work is not to declare decoder superiority, but to demonstrate the usefulness of an alternative representation, which has received comparatively limited systematic evaluation in AES.

Beyond differences between model families, we observe that both representation construction and downstream modeling play an important role. Mean pooling consistently decrease performance over CLS-based representations across both SVM and LightGBM, indicating that the method used to derive document-level embeddings has a measurable impact on downstream effectiveness. Furthermore, while both regressors exhibit similar relative trends,

Model	AES2			ELG		
	QWK	Accuracy	F1	QWK	Accuracy	F1
HC + BERT	0.77 (0.01)	0.62 (0.01)	0.61 (0.01)	0.87 (0.03)	0.67 (0.05)	0.66 (0.05)
HC + BERT Large	0.77 (0.01)	0.62 (0.01)	0.60 (0.01)	0.86 (0.03)	0.64 (0.05)	0.63 (0.06)
HC + ModernBERT	0.76 (0.01)	0.61 (0.01)	0.60 (0.01)	0.86 (0.03)	0.65 (0.06)	0.64 (0.06)
HC + DeBERTa	0.77 (0.01)	0.62 (0.01)	0.61 (0.01)	0.89 (0.03)	0.69 (0.07)	0.68 (0.07)
HC + DeBERTa Large	0.77 (0.01)	0.61 (0.01)	0.60 (0.01)	0.87 (0.03)	0.65 (0.07)	0.64 (0.07)
HC + Mistral 7B	0.81 (0.01)	0.66 (0.01)	0.66 (0.01)	0.90 (0.02)	0.72 (0.05)	0.71 (0.05)
HC + QWEN3 0.6B	0.80 (0.01)	0.65 (0.02)	0.65 (0.02)	0.91 (0.02)	0.74 (0.04)	0.74 (0.04)
HC + QWEN3 1.7B	0.81 (0.01)	0.66 (0.01)	0.65 (0.01)	0.90 (0.02)	0.71 (0.06)	0.70 (0.06)
HC + QWEN3 4B	0.81 (0.01)	0.66 (0.01)	0.65 (0.02)	0.89 (0.02)	0.70 (0.04)	0.69 (0.04)
HC + QWEN3 8B	0.81 (0.01)	0.66 (0.01)	0.65 (0.01)	0.90 (0.02)	0.72 (0.03)	0.71 (0.04)
HC + QWEN3 14B	0.81 (0.01)	0.66 (0.01)	0.65 (0.02)	0.91 (0.02)	0.72 (0.03)	0.71 (0.04)
HC + Llama 1B	0.81 (0.01)	0.65 (0.01)	0.65 (0.01)	0.90 (0.02)	0.72 (0.04)	0.71 (0.04)
HC + Llama 3B	0.81 (0.01)	0.66 (0.01)	0.65 (0.01)	0.91 (0.02)	0.75 (0.04)	0.74 (0.04)
HC + Deepseek 1.5B	0.80 (0.01)	0.65 (0.01)	0.64 (0.01)	0.88 (0.03)	0.69 (0.03)	0.68 (0.04)
HC + Deepseek 7B	0.81 (0.01)	0.65 (0.01)	0.65 (0.02)	0.88 (0.02)	0.69 (0.06)	0.68 (0.06)

Table 6: Hybrid configuration (HC + embeddings). Mean and standard deviation across folds.

LightGBM consistently outperforms linear SVM, reflecting the presence of non-linear structure in the embedding space. Taken together, these findings reinforce that both representation extraction and downstream modeling choices influence how effectively transformer embeddings can be leveraged for AES.

The absence of consistent scaling effects contrasts with broader claims about generative model scaling (Kaplan et al., 2020). While scaling laws hold for language modeling objectives, their impact in the embedding space appears less straightforward. In our findings for AES, moderate-sized decoder models achieve competitive performance without clear monotonic gains from additional parameters. The findings suggest that scaling effects observed in generative language modeling do not necessarily translate into proportional gains in representation-level transfer for structured assessment tasks.

Hybrid AES architectures have consistently shown gains from integrating linguistic features (Dasgupta et al., 2018; Uto et al., 2020). While linguistic features yield consistent improvements across architectures, they do not alter the relative ranking between encoder and decoder backbones. The implication is not that linguistic features are dispensable, but that representation learning can be enriched with linguistic information to improve its representation in some contexts.

It is important to delimit the scope of inference. The study does not isolate the effect of training objective, architecture, or pre-training data composition. Encoder and decoder families differ along multiple axes, and the observed differences should be interpreted at the level of model families rather than single causal mechanisms. Consequently, the conclusions are bounded to English-language corpora and the evaluated regression task. Alternative downstream models or multilingual settings may alter relative performance.

6. Conclusion

This paper presented a controlled comparison between encoder-based and decoder-based transformer embeddings for Automated Essay Scoring in a prompt-agnostic setting. By evaluating frozen representations within a fixed regression framework across two corpora (AES2 and ELG), we isolated representation-level differences from adaptation effects.

Our findings answer the research questions as follows. First, decoder embeddings consistently match or outperform encoder embeddings under identical downstream conditions. Second, scaling effects within decoder families are modest and not systematically monotonic. Third, integrating linguistic features yields consistent gains but does not alter the relative ranking between encoder and decoder backbones. Finally, the strongest frozen decoder embeddings achieve performance competitive with a fine-tuned BERT baseline.

The main contribution of this work is a systematic, representation-focused evaluation of transformer model families for AES, disentangled from prompt-

ing and fine-tuning strategies. The results indicate that decoder-based embeddings constitute a viable and underexplored backbone for assessment tasks.

Future work should investigate multilingual settings, alternative downstream architectures, and controlled comparisons using matched model families to further clarify the sources of representational differences.

Limitations

While two datasets were used, both are English-language corpora. The generalizability of these findings to multilingual AES remains to be investigated. Also, the study focuses on frozen embeddings. Alternative downstream architectures may interact differently with transformer representations.

Plain Summary

Large Language Models are now widely used for automated essay scoring, usually either by fine-tuning encoder models or by using decoder models with prompts. However, it is still unclear how good decoder models are when used simply as fixed feature extractors (without prompting or training). In this paper, we compare encoder and decoder models in a controlled way, using them only to create essay representations. We test these representations with standard regression models on two English essay datasets. We also look at whether model size matters and whether adding linguistic features (like grammar or vocabulary measures) improves results. We find that decoder-based representations consistently perform better than encoder-based ones when used on their own, and this holds across different types of essay scoring tasks. Increasing model size only leads to small improvements. When we combine model representations with linguistic features, performance improves further. Importantly, the best decoder-based representations perform about as well as a fine-tuned BERT model, even without additional training. Overall, the results show that how models represent text (their embeddings) plays a key role in essay scoring, and that decoder models are a strong and underused option for this task.

Bibliographical References

Y. Attali and J. Burstein. 2006. Automated essay scoring with e-rater[r] v.2. *The Journal of Technology, Learning, and Assessment*, 4(3).

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

M. Breuker. 2023. Cefr labelling and assessment services. In G. Rehm, editor, *European Language Grid*, Cognitive Technologies. Springer, Cham.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Scott Crossley, Perpetual Baffour, Jules King, Lauryn Burleigh, Walter Reade, and Maggie Demkin. Learning agency lab-automated essay scoring 2.0, kaggle (2024). *URL https://kaggle. com/competitions/learning-agency-lab-automated-essay-scoring-2*.

T. Dasgupta, A. Naskar, L. Dey, and R. Saha. 2018. Augmenting textual qualitative features in deep convolution recurrent neural network for automatic essay scoring. In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 93–102.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 2018. Bert: pre-training of deep bidirectional transformers for language understanding.

P. W. Foltz, D. Laham, and T. K. Landauer. 1999. The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, 1(2):939–944.

L. Hansen, L. R. Olsen, and K. Enevoldsen. 2023. Textdescriptives: A python package for calculating a large variety of metrics from text.

P. He, J. Gao, and W. Chen. 2021. Debertav3: improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.

P. He, X. Liu, J. Gao, and W. Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention.

J. M. Imperial, A. Barayan, R. Stodden, R. Wilkens, R. M. Sanchez, L. Gao, and H. T. Madabushi. 2025. Universalcefr: Enabling open multilingual research on language proficiency assessment. *arXiv preprint arXiv:2506.01419*.

A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. 2023. Mistral 7b.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Beata Beigman Klebanov and Nitin Madnani. 2022. *Automated essay scoring*. Springer Nature.

P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. Roberta: a robustly optimized bert pretraining approach.

Llama Team, AI @ Meta. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

E. Mayfield and A. W. Black. 2020. Should you fine-tune bert for automated essay scoring? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 151–162.

A. Mizumoto and M. Eguchi. 2023. Exploring the potential of using an ai language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2):100050.

E. B. Page. 1966. The imminence of... grading essays by computer. *The Phi Delta Kappan*, 47(5):238–243.

L. Perelman. 2014. When "the state of the art" is counting words. *Assessing Writing*, 21:104–111.

A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. 2018. Improving language understanding by generative pre-training.

P. U. Rodriguez, A. Jafari, and C. M. Ormerod. 2019. Language models and automated essay scoring.

A. Rogers, O. Kovaleva, and A. Rumshisky. 2020. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

M. D. Shermis and J. Burstein. 2013. *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, 1st edition. Routledge.

M. Stahl, L. Biermann, A. Nehring, and H. Wachsmuth. 2024. Exploring llm prompting strategies for joint essay scoring and feedback generation. *arXiv preprint arXiv:2404.15845*.

K. Taghipour and H. T. Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1882–1891.

M. Uto, Y. Xie, and M. Ueno. 2020. Neural automated essay scoring incorporating handcrafted features. In *Proceedings of the 28th international conference on computational linguistics*, pages 6077–6088.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. 2017. Attention is all you need.

Y. Wang, Z. Wei, Y. Zhou, and X. J. Huang. 2018. Automatic essay scoring incorporating rating schema via reinforcement learning. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 791–797.

T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, and A. M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

K. P. Yancey, G. Laflair, A. Verardi, and J. Burstein. 2023. Rating short l2 essays on the cefr scale with gpt-4. In *Proceedings of the 18th workshop on innovative use of NLP for building educational applications (BEA 2023)*, pages 576–584.

Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le. 2019. Xlnet: generalized autoregressive pretraining for language understanding.

T. Zesch, M. Wojatzki, and D. Scholten-Akoun. 2015. Task-independent features for automated essay grading. In *Proceedings of the tenth workshop on innovative use of NLP for building educational applications*, pages 224–232.

Proficiency-Controlled Text Simplification in European Portuguese: A Preliminary Study using Prompting Approaches

Eugénio Ribeiro[1,2], **David Antunes**[1], **Nuno Mamede**[1], **Jorge Baptista**[1,3]
[1]INESC-ID Lisboa, Portugal
[2] Instituto Universitário de Lisboa (ISCTE-IUL), Portugal
[3] Faculdade de Ciências Humanas e Sociais, Universidade do Algarve, Portugal
{eugenio.ribeiro,david.f.l.antunes,nuno.mamede,jorge.baptista}@inesc-id.pt

Abstract

This paper presents a preliminary study on proficiency-controlled text simplification in European Portuguese using multiple prompting strategies. We focus on the iRead4Skills dataset, which defines four complexity levels targeted at adult native speakers with low literacy. Specifically, we simplify 40 texts from the highest complexity level into three easier levels (*plain*, *easy*, and *very easy*), corresponding approximately to Common European Framework of Reference for Languages (CEFR) levels B1, A2, and A1. We evaluate zero-shot and few-shot prompting configurations, exploring the impact of CEFR anchoring, explicit meaning-preservation instructions, and example-based guidance. Automatic evaluation relies on a fine-tuned proficiency classifier and semantic similarity metrics, including BERTScore and document embeddings. The results show that while exact target-level accuracy remains below 40%, target-or-below accuracy reaches up to 61.39%, indicating that the model generally simplifies texts but struggles to consistently match precise proficiency targets. Human evaluation confirms the overall trends observed automatically, while highlighting the subjectivity inherent to proficiency assessment and meaning preservation. Our findings suggest that prompt engineering alone is insufficient for robust proficiency control in European Portuguese, motivating future work on model adaptation and improved evaluation protocols.

Keywords: Text Simplification, Proficiency, European Portuguese

1. Introduction

Text simplification aims to make written content more accessible while preserving its essential meaning. However, unconstrained simplification does not necessarily ensure accessibility, as it may fail to address the specific needs of target readers. For this reason, the research community has increasingly moved toward targeted simplification for social good (Stajner, 2021).

In recent years, Large Language Models (LLMs) have demonstrated strong performance across a wide range of generative tasks, including summarization and rewriting, which are closely related to text simplification (Li et al., 2025; Zhang et al., 2026). Nevertheless, controlling the linguistic complexity of generated texts remains challenging, particularly when targeting predefined proficiency levels (Alva-Manchego et al., 2025).

Readability-controlled text simplification requires systems to adapt outputs to specific levels of linguistic competence. While this task has been extensively explored for English (e.g. Scarton and Specia, 2018; Malik et al., 2024; Alva-Manchego et al., 2025), research on proficiency-controlled simplification for European Portuguese remains limited. Moreover, it is unclear to what extent prompting strategies can guide LLMs toward precise proficiency targets without resorting to computationally expensive ensemble approaches, multiple candidate generation, or model fine-tuning.

In this paper, we present a preliminary study on proficiency-controlled text simplification in European Portuguese using GPT-5-nano OpenAI (2025), a compact large language model with good summarization performance. We focus on the iRead4Skills dataset (Pintard et al., 2024), which defines four complexity levels designed for adult native speakers with low literacy skills. Specifically, we simplify texts from the highest complexity level into three easier levels (*plain*, *easy*, and *very easy*), approximately corresponding to Common European Framework of Reference for Languages (CEFR) levels B1, A2, and A1 (Council of Europe, 2001).

Our objective is to systematically examine how different prompting configurations influence proficiency alignment in European Portuguese text simplification. We compare multiple strategies, including CEFR anchoring of target levels, explicit meaning-preservation instructions, and few-shot examples retrieved via semantic similarity. Through this comparison, we seek to better understand the contribution of prompt design to readability control. Evaluation is conducted automatically using a fine-tuned textual complexity classifier and semantic similarity metrics, and complemented with human assessment. The results provide empirical insights into the capabilities and limitations of current LLMs in this setting, while the simplified texts generated during human evaluation form a small parallel resource that may support future research.

Proceedings of the Joint Workshop on Readability and Text Simplification (READIxTSAR) @ LREC 2026, pages 151–163
11 May 2026.

In the remainder of this paper, we start by providing an overview of related work in Section 2. Then, Section 3 describes the experimental setup, including the dataset, prompting strategies, and evaluation methodology. Section 4 presents the experimental results. Finally, Section 5 summarizes the contributions of this study and provides pointers for future research.

2. Related Work

Overall, text Simplification aims to transform a text into a linguistically simpler version while preserving its original meaning and discourse function. Traditionally, text simplification has been divided into lexical simplification (substitution of complex words), syntactic simplification (sentence splitting, reordering, structural reduction), and, more recently, document-level and controllable simplification. Evaluation typically combines automatic metrics such as BLEU (Papineni et al., 2002) and SARI (Xu et al., 2016), semantic similarity measures such as BERTScore (Zhang et al., 2020), and human judgments quantified through agreement measures such as Cohen's κ (Cohen, 1960). Below, we provide an overview on proficiency-controlled text simplification. However, considering that the task is underexplored in Portuguese, especially in its European variety, we also provide an overview on generic text simplification in Portuguese.

2.1. Proficiency-Controlled Text Simplification

Early work on readability-controlled simplification relied on professionally curated parallel corpora. The Newsela corpus (Xu et al., 2015), which contains news articles rewritten by professional editors to match multiple grade levels, has been particularly influential. Larger automatically aligned resources, such as Newsela-Auto (Jiang et al., 2020), extended this paradigm by generating proficiency-aligned sentence pairs at scale. Beyond sentence-level simplification, Uchida et al. (2018) introduced a dataset for CEFR-controlled lexical simplification, highlighting the importance of word-level proficiency distinctions.

Building on these datasets, several approaches incorporated explicit control mechanisms into neural models. Scarton and Specia (2018) applied a Machine Translation (MT)-inspired sequence-to-sequence architecture to targeted simplification using proficiency-annotated parallel corpora. Subsequent work explored control-token-based conditioning strategies. Nishihara et al. (2019) introduced target-grade level tokens at the sentence level and further incorporated lexical control by weighting the training loss according to word distributions associated with specific grade levels. Similarly, Zetsu et al. (2022) modeled simplification as a sequence of edit operations conditioned on a target level, generating lexical constraints to encourage or discourage specific word choices. Extending this line of research, Agrawal and Carpuat (2023) conducted a systematic analysis of control tokens, examining how different token configurations affect simplification quality and proposing a method to predict low-level control signals at inference time based on the source text and desired target grade.

Motivated by its impact on the performance of instruction-tuned LLMs, several studies have incorporated Reinforcement Learning (RL) in proficiency-controlled text simplification to improve the alignment between the generated outputs and the target levels. For instance, Yanamoto et al. (2022) combined a sequence-to-sequence architecture with deep RL, using a reward function based on the deviation between the generated sentence's predicted difficulty and the intended target level. In this context, readability control has also been explored in adjacent generation tasks, such as summarization, which can be seen as a kind of simplification when the target level is less complex than the original. In particular, Ribeiro et al. (2023) addressed readability-controlled summarization by fine-tuning a sequence-to-sequence model with instruction prompting and RL to target specific Flesch Reading Ease scores (Kincaid et al., 1975). Additionally, they explored advanced decoding strategies based on lookahead search, improving performance but significantly increasing the computational cost.

With the rise of LLMs, recent work has increasingly relied on prompting rather than fine-tuning. Imperial and Tayyar Madabushi (2023) evaluated zero-shot prompting strategies for proficiency-controlled simplification using both Flesch-Kincaid defined grades (Kincaid et al., 1975) and CEFR levels as target and relying on fine-tuned proficiency classifiers for automatic evaluation. Similarly, Farajidizaji et al. (2024) investigated zero-shot readability control, comparing single-step and iterative rewriting strategies and observing that two-step approaches can improve alignment. (Malik et al., 2024) systematically studied the impact of prompt design, showing that explicit CEFR descriptions and few-shot examples improve performance, and that fine-tuning and RL further enhance controllability. Focusing on sentence-level simplification, (Barayan et al., 2025) demonstrated that combining level descriptions with multiple examples of each target level yields the strongest results, while also emphasizing the difficulty of handling large proficiency gaps and the limitations of automatic evaluation metrics.

The growing interest in readability control culminated in the TSAR 2025 Shared Task on Readability-Controlled Text Simplification (Alva-

Manchego et al., 2025). The task, focusing on the simplification of 100 texts with human-generated references to multiple CEFR target levels, attracted 20 participating teams. The evaluation combined automatic CEFR-level classification using models trained on UniversalCEFR (Imperial et al., 2025) and meaning preservation assessment by computing MeaningBERT (Beauchemin et al., 2023) between the generated texts and both the source and reference texts. Most high-performing systems relied on LLMs, often incorporating iterative refinement, ensemble strategies, LLM-as-a-judge frameworks, or external data. Commercial models generally outperformed open-weights models, potentially due to sheer number of parameters, and ensemble approaches frequently surpassed single-model systems, albeit at substantial computational cost. The top-ranked system, by the EhiMeNLP team (Miyata et al., 2025), combined multiple LLMs and several prompting strategies, while other leading teams employed candidate selection via Minimum Bayes Risk decoding (e.g. Hayakawa et al., 2025) or iterative self-refinement guided by CEFR feedback (e.g. Shimada et al., 2025). Overall, the shared task highlighted both the promise of LLM-based readability control and the difficulty of achieving precise, cost-effective level alignment.

2.2. Text Simplification in Portuguese

While English benefits from large parallel corpora, Portuguese has historically faced resource scarcity, which has shaped the development of its simplification approaches.

Early research on Portuguese text simplification was predominantly rule-based and readability-oriented. The PorSimples project (Aluísio et al., 2008a,b; Cândido Junior et al., 2009a,b) focused on syntactic transformation rules and readability assessment, leading to aligned corpora such as PorSimplesSent (Leal et al., 2018). Rule-based syntactic simplification (Cândido Junior et al., 2011) achieved promising results in controlled settings, though coverage remained limited compared to human rewrites. In parallel, studies on readability modeling and textual complexity features (Amancio et al., 2011) contributed to the computational characterization of linguistic difficulty in Brazilian Portuguese.

Statistical MT (SMT) paradigms were subsequently adapted to simplification. Specia (2010) framed text simplification as monolingual translation, demonstrating that SMT could capture lexical operations but often produced conservative outputs. Neural MT approaches later improved fluency and adequacy when parallel data was available (de Lima et al., 2021), and sentence compression strategies were also explored (Nóbrega et al., 2020). These works marked a transition toward data-driven simplification methods, though evaluation setups varied considerably.

Meanwhile, at the lexical level, substitution lexicons were still developed (Wilkens et al., 2017) and research focused on automatically identifying psycholinguistic properties of words for subsequent use in text simplification (Santos et al., 2017). Additionally, hybrid linguistic-statistical simplification architectures for Ibero-Romance languages (Ferrés et al., 2017) demonstrated strong morphological generation capabilities, though word-sense disambiguation remained a limiting factor.

More recently, large pretrained language models and unsupervised style-transfer techniques have gained prominence. Scalercio et al. (2024) trained PT-T5 (Carmo et al., 2020) using paraphrase triplets mined from Common Crawl, mitigating the scarcity of parallel corpora. This led to improvements in SARI over MUSS (Martin et al., 2022), an unsupervised multilingual simplification method, and an LLM baseline on the PorSimplesSent benchmark, while maintaining strong semantic preservation. However, it lost to the LLM on the MUSEUM-PT (Finatto and Tcacenco, 2021) benchmark, as well as on a translation of ASSET (Alva-Manchego et al., 2020).

At the lexical level, North et al. (2024) introduced MultiLS-PT, a multi-genre dataset for lexical simplification, and explored the use of several multilingual and Portuguese-specific models for lexical complexity prediction and substitute generation. While the top performance on the former was achieved using a fine-tuned version of BERTimbau (Souza et al., 2020), the latter was dominated by LLMs.

Additional work has explored different simplification approaches for domain-specific rewriting, with special focus on the legal domain (e.g. Alves et al., 2023; Pereira et al., 2024).

Despite clear progress, three recurring challenges remain: (1) limited high-quality parallel and multi-reference corpora compared to English; (2) domain generalization, as models trained on translated or narrow-domain corpora may degrade on authentic data; and (3) user-centered adaptation, particularly for audiences with low literacy levels or specific accessibility needs. While Portuguese text simplification has evolved from rule-based syntactic rewriting to neural and LLM-based paradigms, systematic studies on fine-grained proficiency control remain scarce, especially for European Portuguese. Furthermore, robust evaluation frameworks and broader-coverage resources remain essential for reliable real-world deployment.

3. Experimental Setup

In this section, we describe the experimental setup adopted to evaluate proficiency-controlled

text simplification in European Portuguese, including the dataset, prompting strategies, and evaluation methodology.

3.1. Dataset

The iRead4Skills corpus (Pintard et al., 2024) consists of texts in three languages—French, Portuguese, and Spanish—, classified by human experts into four complexity levels, roughly corresponding to CEFR (Council of Europe, 2001) levels, but targeted at adult native speakers with low literacy (Monteiro et al., 2023):

Very Easy: Texts that are fully or almost fully understood by everyone, including people with very low schooling (i.e., that did not finish the primary school) and almost no reading experience. (CEFR Level A1)

Easy: Texts that are fully or almost fully understood by people with low schooling (i.e., that completed the primary school but do not have more than the 9th year) and have poor reading experience. (CEFR Level A2)

Plain: Texts that are understood the first time they are read by people that completed the 9th year and have a functional-to-average reading experience. (CEFR Level B1)

More Complex: Texts with a higher complexity than that defined by the previous levels. (CEFR Levels B2 and above)

In this study, we focus on the Portuguese data in the dataset (Reis et al., 2024). Specifically, as it is a preliminary study, we take the 40 texts of the more complex level in the test set defined by Ribeiro et al. (2025) and explore their simplification to the three easier levels. Additionally, we use the training set as a source of examples in the few-shot setting.

3.2. Prompting

Considering this is a preliminary study, for speed and cost purposes, we rely solely on the GPT-5 nano model (OpenAI, 2025), which is claimed to be good for summarization tasks. Still, we assess the impact of different prompt components in both zero-shot and few-shot settings. The wording for each prompt component is based on a small set of pilot experiments, exploring a few alternative phrasings and retaining those that performed most consistently across examples. Figure 1 shows the full simplification prompt template.

3.2.1. Zero-Shot Setting

The base prompt (unshaded blocks in Figure 1) fills four main purposes:

1. Stating the context: simplification targeted at adult speakers with a given proficiency level;
2. Instructing the system to reply with the simplified text only;
3. Describing the target proficiency level using the Portuguese version of the descriptions in Section 3.1, without the information about CEFR level approximation;
4. Providing the text to simplify.

Additionally, we explore the impact of two factors:

1. Explicitly instructing the system to keep the essential information and original meaning (*KeepInfo*);
2. Pairing the proficiency level description with its CEFR approximation (*CEFR*).

While the former aims to assess whether the LLM intrinsically tends to diverge from the original content, the latter aims to assess whether the model can leverage intrinsic knowledge regarding the CEFR to produce a more appropriate simplification to the target level.

3.2.2. Few-Shot Setting

In the few-shot setting, we explore both 1-shot and 3-shot approaches. The examples are selected from the training subset of the iRead4Skills dataset. We explore example selection from two pools: all the texts of the target level and the texts of the target level of the same genre as the text to be simplified. To select the examples, we rely on semantic search as provided by the Sentence Transformers library (Reimers and Gurevych, 2019), that is, we select the closest examples to the text to be simplified. To generate document embeddings, we use the Serafim model with 900M parameters tuned for Information Retrieval (IR) (Gomes et al., 2025).

3.3. Evaluation Methodology

We split our evaluation procedure in two steps: first, we perform automatic evaluation to assess the differences between the multiple experimental conditions and select the best one and, then, we rely on human annotators to assess the appropriateness of the generated simplifications.

3.3.1. Automatic Evaluation

To assess whether the simplified texts match the target proficiency level, we rely on the fine-tuned model developed by Ribeiro et al. (2025) and used in the context of the iRead4Skills project (Aissa et al., 2025).This model was trained on the

Simplification Prompt Template

Simplifica o seguinte texto, de modo a que ele seja compreensível por adultos falantes nativos de português europeu com o nível de proficiência indicado.
Simplify the following text so that it is understandable to adult native speakers of European Portuguese with the indicated proficiency level.

[1]Ao simplificar o texto, mantém a informação essencial e o significado original.
When simplifying the text, keep the essential information and the original meaning.

[3]Em conjunto com cada texto, vais receber alguns exemplos do nível indicado para ajudar a guiar a simplificação.
Along with each text, you will receive examples of the target level to guide the simplification.

Responde apenas com o texto simplificado, sem introduções ou explicações adicionais.
Respond only with the simplified text, without introductions or additional explanations.

Nível de proficiência alvo: **[LEVEL DESCRIPTION]** [2](Nível **[CEFR LEVEL]** do CEFR)
Target proficiency level: ***[LEVEL DESCRIPTION]*** *(CEFR Level* ***[CEFR LEVEL]****)*

[3]Exemplo do nível alvo: **[EXAMPLE]**
Example of the target level: ***[EXAMPLE]***

Texto a simplificar: **[TEXT]**
Text to simplify: ***[TEXT]***

[1] Included only in the *KeepInfo* configuration.
[2] Included only in the *CEFR* configuration.
[3] Included only in the few-shot setting. The second block is repeated for every example.

Figure 1: Full simplification prompt template. Shaded segments indicate components that vary across experimental configurations.

iRead4Skills dataset and classifies texts according to the proficiency level required to understand them, using the four-level scale.

In the context of the simplification task, a text with a required proficiency level below the target is still considered appropriate. Accordingly, we use target-or-below (ToB) accuracy as the main evaluation metric. To enable a more in-depth analysis, we also adopt some of the most common evaluation metrics used in previous studies on automatic readability level classification: accuracy, adjacent accuracy, and the macro F_1 score. Adjacent accuracy allows for deviations of one level from the target, enabling the identification of settings with more extreme deviations. Comparing macro F_1 and accuracy scores provides further insight into whether simplifying to certain target levels is more difficult than others.

Another crucial aspect for the simplification task is meaning preservation. We assess this using the BERTScore F_1 (Zhang et al., 2020), scaled with the Portuguese baseline of the large XLM-RoBERTa model (Conneau et al., 2020), as well as document similarity based on the document embeddings generated by the Serafim model.

Considering the non-determinism nature of LLMs, we perform three runs of each experiment. The results reported in Section 4.1 correspond to the average and standard deviation across the three runs. All of the metrics are reported in percentage form.

3.3.2. Human Evaluation

The simplified texts produced by the best-performing run, as identified through automatic

evaluation, were selected for subsequent human assessment. The proficiency level required to understand each simplified text was annotated according to the four iRead4Skills levels. Meaning preservation relative to the source text was evaluated using a 4-point Likert scale, where 0 denotes that none of the essential information was retained and the meaning was substantially altered, and 3 denotes that all essential information was preserved and the meaning remained unchanged.

Three experts in linguistics and education annotated 20% of the simplified texts to estimate inter-annotator agreement using Krippendorff's α (Hayes and Krippendorff, 2007). The remaining texts were annotated by at least one expert. For each text, the final label corresponds to the average of the assigned annotations. To promote transparency and support future research, we plan to publicly release the annotated dataset.

The performance of both the simplification approach and the readability classifier can be evaluated against the human annotations using the same classification metrics adopted in the automatic evaluation.

4. Results

In this section, we present and analyze the results of the experimental evaluation. We first report the outcomes of the automatic evaluation across the different prompting configurations, comparing their performance in terms of proficiency alignment and meaning preservation. We then examine the findings of the human evaluation to assess the extent to which automatic metrics reflect human judgments and to provide a more nuanced interpretation of the model's behavior.

4.1. Automatic Evaluation

Table 1 shows the results of the automatic evaluation. We can see that the base prompt only achieves 35.28% accuracy in terms of the target level and that it stays under 40% for every configuration. Additionally, on average, the macro F_1 stays under 30% for every configuration. This suggests that the model struggles to generate texts that align with a specific proficiency level. However, two major aspects must be taken into consideration. First, readability assessment has proven to be a very subjective task with low agreement even among humans (e.g. Branco et al., 2014; Curto, 2014; Ribeiro et al., 2024, 2025) and with the automatic classifier achieving around 52% accuracy on the iRead4Skills dataset (Ribeiro et al., 2025). This is consistent with the significantly higher adjacent accuracy (Adj. Acc.). Second, and most importantly, the prompt instructs the model to generate a text that is understandable by speakers with the indicated proficiency level. Thus, generating a text with a lower proficiency requirement still fulfills the task. This is captured by our main metric, target-or-below accuracy (ToB Acc.), which is significantly higher than accuracy for every configuration, with a lowest average score of 56.11% for the base prompt. This indicates that the model more frequently oversimplifies than undersimplifies.

Looking into the meaning preservation metrics, an average BERTScore around 50% suggests that there is some information loss or meaning change. However, that is not surprising in a simplification task, as parts of the original text are discarded or written in simpler terms. Furthermore, the document similarity is above 90% for every configuration, which suggests that there are no major divergences in meaning.

Going into further detail regarding the differences between the multiple prompt configurations, in the first block of Table 1, which corresponds to the zero-shot setting, we can see that both additional components affect the performance in different ways. As expected, explicitly instructing the model to keep the essential information and the original meaning improves the meaning preservation metrics while having minimal impact on the target proficiency metrics. On the other hand, adding the CEFR level approximation to the target level description improves target-or-below accuracy, but decreases accuracy and the meaning preservation metrics. The former suggests that the model can leverage prior knowledge of CEFR descriptors in the targeted simplification process. However, the decrease in accuracy may indicate a partial mismatch between the iRead4Skills levels and their CEFR approximations, with the former being slightly harder. Finally, the reduction in terms of the meaning preservation metrics is expected, considering that the model is introducing parts of its intrinsic knowledge regarding the CEFR in the simplification.

Overall, combining both additional components leads to the best performance in terms of target-or-below accuracy in the zero-shot setting. Thus, we used it as the starting point for the few-shot setting. Looking into the results in the second block of Table 1, we can see the highest scores in terms of all target proficiency metrics, except for macro F_1. The latter is an anomaly, caused by an outlier run in the zero-shot setting which achieved a significantly higher score. On the other hand, BERTScore is further impacted, potentially by the introduction of words, expressions, or structures present in the examples that do not appear in the original texts. Still, the document similarity stays in line with the zero-shot setting.

Looking into the results in further detail, we can see that using multiple examples can be harmful,

Prompt	ToB Acc.	Accuracy	Adj. Acc.	Macro F_1	BERTScore	Similarity
Base	56.11±1.42	35.28±1.04	84.44±1.57	24.70±1.00	53.71±0.68	93.42±0.25
KeepInfo	56.39±1.04	35.56±1.71	85.28±1.71	24.67±1.32	**54.89±0.27**	**94.03±0.21**
CEFR	59.44±1.04	34.17±0.68	89.44±0.79	24.45±0.75	49.52±0.35	92.33±0.16
CEFR+KeepInfo	60.28±1.04	36.67±0.68	88.89±1.04	**28.44±4.71**	50.64±0.71	92.45±0.35
1-shot	**61.39±0.39**	**38.89±0.79**	89.72±1.04	28.07±0.38	49.02±0.64	92.69±0.58
1-shot (Genre)	59.44±1.04	36.39±1.04	**90.28±0.39**	25.66±1.55	49.13±0.49	92.68±0.55
3-shot	60.83±1.18	34.72±1.04	89.44±2.08	25.47±0.48	47.44±0.49	92.41±0.55
3-shot (Genre)	60.00±1.36	37.50±2.04	89.72±1.04	27.02±1.78	47.43±0.62	92.41±0.24

Table 1: Automatic evaluation results. The few-shot experiments build on the CEFR+KeepInfo setting.

as the top performance is achieved in the 1-shot setting. Furthermore, restricting the examples to the same genre as the original text seems to be harmful as well. While the former can be explained by the long-context degradation phenomenon in LLMs, the latter suggests that the restricted pool of examples may be too small to provide relevant examples that cover similar subjects. Overall, a target-or-below accuracy of 61.39% shows that the GPT-5 nano model struggles to consistently provide appropriate simplifications for the target proficiency level. Still, the automatic classifier may be overestimating the level. Thus, in the next section, we discuss the results of the human evaluation.

4.2. Human Evaluation

We selected for human evaluation the texts generated by the top-performing run in terms of target-or-below accuracy. Inter-annotator agreement among the three experts on the required proficiency level was moderate (Krippendorff's $\alpha = 0.50$), and notably higher than the low agreement levels previously reported for Portuguese readability assessment (Branco et al., 2014; Curto, 2014; Ribeiro et al., 2024). This confirms that the description of the proficiency levels and their CEFR approximation remains open to subjective interpretation. Agreement on meaning preservation was considerably lower ($\alpha = 0.01$), indicating substantial divergence in what annotators considered essential information. A clear example emerged in the simplification of cooking recipes: one annotator disregarded missing ingredients, whereas others assigned low preservation scores in such cases. This illustrates the inherent difficulty of defining essential information in simplification tasks without additional contextual criteria. Nevertheless, the average meaning preservation score was 2.58 out of 3, consistent with the high document similarity observed in the automatic evaluation.

Table 2 reports the target proficiency metrics obtained when comparing the human annotations with both the intended target levels and the automatically predicted levels. The target-or-below accuracy reaches 62.50%, which is 1 percentage point higher than that automatically computed for the same run. In contrast, accuracy and macro F_1 decrease by approximately 3 percentage points, while adjacent accuracy remains unchanged. These results suggest that the automatic classifier exhibits a slight tendency to overestimate the proficiency level required to understand the simplified texts. When evaluated against the human annotations, the classifier achieves 42.50% exact accuracy but 95% adjacent accuracy, further reflecting the inherent subjectivity of the task, as discussed in previous studies and evidenced by the low inter-annotator agreement. A closer inspection of the predictions shows that both the human annotators and the classifier assign the *easy* level to 57% of the texts. However, the classifier frequently labels *very easy* texts as *easy*, and several *easy* texts as *plain*.

Focusing on the targeted simplification task, two main conclusions emerge. First, despite some discrepancies, automatic evaluation appears to be a reliable proxy for human assessment in terms of overall results, as reflected in the similar target-or-below accuracy and the consistency between document similarity scores and human judgments of meaning preservation. Second, in line with the automatic results, GPT-5 nano struggles to consistently generate simplifications that precisely match the intended proficiency level. Nevertheless, both the automatic classifier and the human annotators identified only three texts as *more complex*, indicating that the model generally reduces textual complexity, albeit not always sufficiently to ensure full comprehensibility for speakers at the target proficiency level.

5. Conclusion

This paper presented a preliminary study on proficiency-controlled text simplification in European Portuguese using GPT-5 nano paired with multiple prompting strategies in both zero-shot and few-shot settings. By simplifying texts from the highest complexity level of the iRead4Skills dataset to three easier levels, we assessed the model's abil-

	ToB Acc.	Accuracy	Adj. Acc.	Macro F_1
Target	62.50	35.00	90.00	25.83
Predicted	-	42.50	95.00	29.99

Table 2: Comparison of the human annotations with the target and predicted levels.

ity to align its outputs with predefined proficiency targets.

Results indicate that, although the model consistently reduces textual complexity, it struggles to reliably match the exact target level. The discrepancy between accuracy and target-or-below accuracy suggests a tendency toward slight oversimplification rather than undersimplification. Few-shot prompting improves proficiency alignment, particularly in the 1-shot setting, while adding CEFR information increases compliance but may introduce minor meaning deviations.

Human evaluation broadly confirms the automatic trends, while revealing substantial subjectivity in both proficiency assessment and meaning preservation. Overall, our findings show that, while it may impact performance, lightweight prompt engineering is not sufficient for precise proficiency control, at least in European Portuguese. Future work should explore the use of LLM with different architectures and prompting paradigms, as well as ensemble approaches with candidate selection. Model adaptation is also an option, but it requires effort towards the creation of curated parallel simplification corpora. Furthermore, and perhaps most importantly, additional efforts should be dedicated to the creation of more comprehensive level definitions and improved evaluation frameworks to support reliable accessibility-oriented simplification.

6. Limitations

This study presents several limitations that should be considered when interpreting the results.

First, the experimental setup is restricted to a single LLM and a relatively small dataset consisting of 40 source texts. Although multiple prompting strategies and three independent runs were used to improve robustness, the findings cannot be generalized to other model families or larger-scale scenarios without further experimentation.

Second, automatic evaluation relies on a fine-tuned proficiency classifier. While the classifier provides a consistent and scalable evaluation framework, its predictions do not fully capture the nuanced criteria used by human annotators when assessing readability. The relatively low inter-annotator agreement observed in the human evaluation further highlights the inherent subjectivity of proficiency assessment.

Third, meaning preservation was assessed using semantic similarity metrics and human judgments on a Likert scale. Although this provides complementary perspectives, both approaches have limitations, and the low inter-annotator agreement observed suggests that meaning preservation remains difficult to evaluate reliably.

Finally, this study focuses exclusively on simplification from the highest complexity level and that was the only constraint on source text selection. As the experimental setup did not include human generation of simplified texts, it is possible that some of the texts cannot be simplified to the lowest complexity levels without losing crucial information. Future work should take this into consideration as well as investigate simplifications from the lower complexity levels to assess whether there are emerging patterns related to the gap between the source and target levels.

7. Ethical Considerations

This work addresses text simplification for adult native speakers with low literacy skills, a population for whom accessibility and clarity are essential. Improving readability has the potential to support inclusion and equitable access to information. However, automated simplification also introduces risks.

First, inaccurate proficiency control may lead to oversimplification or unintended loss of essential information. In contexts such as public communication, health information, or legal texts, even minor omissions may have significant consequences. Our results show that precise level alignment remains challenging, underscoring the need for human oversight in high-stakes applications.

Second, LLMs are trained on large-scale data, which may encode cultural, social, or linguistic biases. These biases can influence lexical choices or framing in simplified outputs, particularly when targeting vulnerable populations. We did not conduct a systematic bias analysis in this study, and this remains an important direction for future work.

Finally, the reliance on automatic evaluation metrics and classifier-based proficiency level prediction introduces additional uncertainty. Although automatic measures correlated with human judgments in our experiments, they should not be considered substitutes for comprehensive human evaluation in real-world deployment scenarios.

8. Lay Summary

Making texts easier to read is important for adults with low reading skills. However, it is not enough to just simplify a text. It should also match the reader's level of proficiency and keep the original meaning.

In this study, we explore how well AI language tools can simplify texts in European Portuguese to different levels of complexity for adult readers. We use a collection of texts from the iRead4Skills project, which defines four levels of complexity for adult learners. We take texts from the most difficult level and rewrite them into three easier versions: plain, easy, and very easy.

We test different ways of guiding the system, including giving examples, using clear instructions, and referring to standard proficiency levels (such as A1, A2, and B1). We then assess the results using computer-based measures and human evaluation.

The results show that the systems are generally able to make texts simpler, but they often fail to match the exact target level of difficulty. This suggests that current methods alone are not enough to fully control text simplification. Further work is needed to improve both the systems and the way we evaluate simplified texts.

9. Acknowledgments

This work was supported by Portuguese national funds through Fundação para a Ciência e a Tecnologia (FCT) under projects UID/50021/2025 (DOI:10.54499/UID/50021/2025) and UID/PRR/50021/2025 (DOI:10.54499/UID/PRR/50021/2025) and by the European Commission (Project: iRead4Skills, Grant number: 1010094837, Topic: HORIZON-CL2-2022-TRANSFORMATIONS-01-07, DOI:10.3030/101094837).

10. Bibliographical References

Sweta Agrawal and Marine Carpuat. 2023. Controlling Pre-trained Language Models for Grade-Specific Text Simplification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 12807–12819.

Wafa Aissa, Raquel Amaro, David Antunes, Thibault Bañeras-Roux, Jorge Baptista, Alejandro Catala, Luís Correia, Thomas François, Marcos Garcia, Mario Izquierdo-Álvarez, Nuno Mamede, Vasco Martins, Miguel Neves, Eugénio Ribeiro, Sandra Rodriguez, and Elodie Vanzereven. 2025. The iRead4Skills Intelligent Complexity Analyzer. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations*, pages 73–84.

Sandra Maria Aluísio, Lucia Specia, Thiago Alexandre Salgueiro Pardo, Erick Galani Maziero, Helena De Medeiros Caseli, and Renata Pontin de Mattos Fortes. 2008a. A Corpus Analysis of Simple Account Texts and the Proposal of Simplification Strategies: First Steps Towards Text Simplification Systems. In *Proceedings of the Annual ACM International Conference on Design of Communication (SIGDOC)*, pages 15–22.

Sandra Maria Aluísio, Lucia Specia, Thiago Alexandre Salgueiro Pardo, Erick Galani Maziero, and Renata Pontin de Mattos Fortes. 2008b. Towards Brazilian Portuguese Automatic Text Simplification Systems. In *Proceedings of ACM Symposium on Document Engineering (DocEng)*, pages 240–248.

Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. 2020. ASSET: A Dataset for Tuning and Evaluation of Sentence Simplification Models with Multiple Rewriting Transformations. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4668–4679.

Fernando Alva-Manchego, Regina Stodden, Joseph Marvin Imperial, Abdullah Barayan, Kai North, and Harish Tayyar Madabushi. 2025. Findings of the TSAR 2025 Shared Task on Readability-Controlled Text Simplification. In *Proceedings of the Workshop on Text Simplification, Accessibility and Readability (TSAR)*, pages 116–130.

Alexandre Alves, Péricles B.C. Miranda, Rafael Ferreira Mello, and André Nascimento. 2023. Automatic Simplification of Legal Texts in Portuguese Using Machine Learning. *Frontiers in Artificial Intelligence and Applications*, 379:281–286.

Marcelo Adriano Amancio, Magali Sanches Duran, and Sandra Maria Aluísio. 2011. Automatic Question Categorization: A new Approach for Text Elaboration. *Procesamiento del Lenguaje Natural*, 46:43–50.

Abdullah Barayan, Jose Camacho-Collados, and Fernando Alva-Manchego. 2025. Analysing Zero-Shot Readability-Controlled Sentence Simplification. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 6762–6781.

David Beauchemin, Horacio Saggion, and Richard Khoury. 2023. MeaningBERT: Assessing Meaning Preservation between Sentences. *Frontiers in Artificial Intelligence*, 6.

António Branco, João Rodrigues, Francisco Costa, João Silva, and Rui Vaz. 2014. Assessing Automatic Text Classification for Interactive Language Learning. In *Proceedings of the International Conference on Information Society (i-Society)*, pages 70–78.

Arnaldo Cândido Junior, Ann Copestake, Lucia Specia, and Sandra Maria Aluísio. 2011. Towards an on-demand Simple Portuguese Wikipedia. *Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, pages 137–147.

Arnaldo Cândido Junior, Matheus de Oliveira, and Sandra Maria Aluísio. 2009a. Simplifica: A Simplified Texts Web Authoring System. In *Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia)*.

Arnaldo Cândido Junior, Erick Maziero, Caroline Gasperin, Thiago Alexandre Salgueiro Pardo, Lucia Specia, and Sandra Maria Aluísio. 2009b. Supporting the Adaptation of Texts for Poor Literacy Readers: A Text Simplification Editor for Brazilian Portuguese. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 34–42.

Diedre Carmo, Marcos Piau, Israel Campiotti, Rodrigo Nogueira, and Roberto Lotufo. 2020. PTT5: Pretraining and Validating the T5 Model on Brazilian Portuguese Data. *Computing Research Repository*, arXiv:2008.09144.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8440–8451.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press.

Pedro Curto. 2014. Classificador de Textos para o Ensino de Português como Segunda Língua. Master's thesis, Instituto Superior Técnico, Universidade de Lisboa.

Tiago B. de Lima, André C. A. Nascimento, George Valença, Pericles Miranda, Rafael Ferreira Mello, and Tapas Si. 2021. Portuguese Neural Text Simplification Using Machine Translation. In *Proceedings of the Brazilian Conference on Intelligent Systems (BRACIS)*, pages 542–556.

Asma Farajidizaji, Vatsal Raina, and Mark Gales. 2024. Is it Possible to Modify Text to a Target Readability Level? An Initial Investigation Using Zero-Shot Large Language Models. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9325–9339.

Daniel Ferrés, Horacio Saggion, and Xavier Gómez Guinovart. 2017. An Adaptable Lexical Simplification Architecture for Major Ibero-Romance Languages. In *Proceedings of the Workshop on Building Linguistically Generalizable NLP Systems*, pages 40–47.

Maria José Bocorny Finatto and Lucas Meireles Tcacenco. 2021. Intralingual Translation, Equivalence Strategies and Textual and Terminological Accessibility. *TradTerm*, 37(1):30–63.

Luís Gomes, António Branco, João Silva, João Rodrigues, and Rodrigo Santos. 2025. Open Sentence Embeddings for Portuguese with the Serafim PT* Encoders Family. In *Proceedings of the EPIA Conference on Artificial Intelligence*, pages 267–279.

Akio Hayakawa, Nouran Khallaf, Horacio Saggion, and Serge Sharoff. 2025. UoL-UPF at TSAR 2025 Shared Task: A Generate-and-Select Approach for Readability-Controlled Text Simplification. In *Proceedings of the Workshop on Text Simplification, Accessibility and Readability (TSAR)*, pages 193–210.

Andrew F. Hayes and Klaus Krippendorff. 2007. Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures*, 1(1):77–89.

Joseph Marvin Imperial, Abdullah Barayan, Regina Stodden, Rodrigo Wilkens, Ricardo Muñoz Sánchez, Lingyun Gao, Melissa Torgbi, Dawn Knight, Gail Forey, Reka R. Jablonkai, Ekaterina Kochmar, Robert Joshua Reynolds, Eugénio Ribeiro, Horacio Saggion, Elena Volodina, Sowmya Vajjala, Thomas François, Fernando Alva-Manchego, and Harish Tayyar Madabushi. 2025. UniversalCEFR: Enabling Open Multilingual Research on Language Proficiency Assessment. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9703–9755.

Joseph Marvin Imperial and Harish Tayyar Madabushi. 2023. Flesch or Fumble? Evaluating Readability Standard Alignment of Instruction-Tuned Language Models. In *Proceedings of the Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 205–223.

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. 2020. Neural CRF Model for Sentence Alignment in Text Simplification. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7943–7960.

J. Peter Kincaid, Robert P. Fishburne Jr, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Technical report, Institute for Simulation and Training, University of Central Florida.

Sidney Evaldo Leal, Magali Sanches Durán, and Sandra Maria Aluísio. 2018. A Nontrivial Sentence Corpus for the Task of Sentence Readability Assessment in Portuguese. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 401–413.

Jiawei Li, Yang Gao, Yizhe Yang, Yu Bai, Xiaofeng Zhou, Yinghao Li, Huashan Sun, Yuhang Liu, Xingpeng Si, Yuhao Ye, Yixiao Wu, Yiguan Lin, Bin Xu, Bowen Ren, Chong Feng, and Heyan Huang. 2025. Fundamental Capabilities and Applications of Large Language Models: A Survey. *ACM Computing Surveys*, 58(2).

Ali Malik, Stephen Mayhew, Christopher Piech, and Klinton Bicknell. 2024. From Tarzan to Tolkien: Controlling the Language Proficiency Level of LLMs for Content Generation. In *Findings of the Association for Computational Linguistics: ACL*, pages 15670–15693.

Louis Raphaël Théo Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. 2022. MUSS: Multilingual Unsupervised Sentence Simplification by Mining Paraphrases. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 1651–1664.

Rina Miyata, Koki Horiguchi, Risa Kondo, Yuki Fujiwara, and Tomoyuki Kajiwara. 2025. EhiMeNLP at TSAR 2025 Shared Task: Candidate Generation via Iterative Simplification and Reranking by Readability and Semantic Similarity. In *Proceedings of the Workshop on Text Simplification, Accessibility and Readability (TSAR 2025)*, pages 217–222.

Ricardo Monteiro, Raquel Amaro, Susana Correia, Alice Pintard, Roser Gauchola, Michell Moutinho, and Xavier Blanco Escoda. 2023. iRead4Skills Complexity Levels. Project Deliverable D3.1, iRead4Skills.

Daiki Nishihara, Tomoyuki Kajiwara, and Yuki Arase. 2019. Controllable Text Simplification with Lexical Constraint Loss. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL): Student Research Workshop*, pages 260–266.

Fernando A. A. Nóbrega, Alipio M. Jorge, Pavel Brazdil, and Thiago A. S. Pardo. 2020. Sentence Compression for Portuguese. In *Proceedings of the International Conference on Computational Processing of Portuguese (PROPOR)*, pages 270–280.

Kai North, Tharindu Ranasinghe, Matthew Shardlow, and Marcos Zampieri. 2024. MultiLS: An End-to-End Lexical Simplification Framework. In *Proceedings of the Workshop on Text Simplification, Accessibility and Readability (TSAR)*, pages 1–11.

OpenAI. 2025. GPT-5 System Card. *Computing Research Repository*, arXiv:2601.03267.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318.

Francielle Vasconcellos Pereira, Ana Paula Rodrigues Feitosa Frazão, and Viviane Pereira Moreira. 2024. Automatic Text Simplification for the Legal Domain in Brazilian Portuguese. In *Proceedings of the Brazilian Conference on Intelligent Systems (BRACIS)*, pages 31–45.

Alice Pintard, Thomas François, Justine Nagant de Deuxchaisnes, Sílvia Barbosa, Maria Leonor Reis, Michell Moutinho, Ricardo Monteiro, Raquel Amaro, Susana Correia, Sandra Rodríguez Rey, Marcos Garcia González, Keran Mu, and Xavier Blanco Escoda. 2024. iRead4Skills Dataset 1: Corpora by Complexity Level for FR, PT and SP. Project Deliverable D3.2, iRead4Skills.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Maria Leonor Reis, Sílvia Barbosa, Michell Moutinho, Ricardo Monteiro, Susana Correia, and Raquel Amaro. 2024. Intelligent Support for Low Literacy Adults: The European Portuguese iRead4Skills Corpus. *International Journal of Emerging Technologies in Learning (iJET)*, 19(8):61–81.

Eugénio Ribeiro, David Antunes, Nuno Mamede, and Jorge Baptista. 2025. Exploring Few-Shot Approaches to Automatic Text Complexity Assessment in European Portuguese. *Journal of the Brazilian Computer Society*, 31(1):690–710.

Eugénio Ribeiro, Nuno Mamede, and Jorge Baptista. 2024. Avaliação Automática do Nível de Complexidade de Textos em Português Europeu. *Linguamática*, 16(2):121–145.

Leonardo F. R. Ribeiro, Mohit Bansal, and Markus Dreyer. 2023. Generating Summaries with Controllable Readability Levels. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 11669–11687.

Leandro Borges dos Santos, Magali Sanches Duran, Nathan Siegle Hartmann, Arnaldo Cândido Junior, Gustavo Henrique Paetzold, and Sandra Maria Aluísio. 2017. A Lightweight Regression Method to Infer Psycholinguistic Properties for Brazilian Portuguese. In *Proceedings of the International Conference on Text, Speech, and Dialogue (TSD)*, pages 281–289.

Arthur Scalercio, Maria Finatto, and Aline Paes. 2024. Enhancing Sentence Simplification in Portuguese: Leveraging Paraphrases, Context, and Linguistic Features. In *Findings of the Association for Computational Linguistics: ACL*, pages 15076–15091.

Carolina Scarton and Lucia Specia. 2018. Learning Simplifications for Specific Target Audiences. In *Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 2: Short Papers, pages 712–718.

Mao Shimada, Kexin Bian, Zhidong Ling, and Mamoru Komachi. 2025. HIT-YOU at TSAR 2025 Shared Task: Leveraging Similarity-Based Few-Shot Prompting, Round-Trip Translation, and Self-Refinement for Readability-Controlled Text Simplification. In *Proceedings of the Workshop on Text Simplification, Accessibility and Readability (TSAR)*, pages 231–241.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In *Proceedings of the Brazilian Conference on Intelligent Systems (BRACIS)*, pages 403–417.

Lucia Specia. 2010. Translating from Complex to Simplified Sentences. In *Proceedings of the International Conference on Computational Processing of Portuguese (PROPOR)*, pages 30–39.

Sanja Stajner. 2021. Automatic Text Simplification for Social Good: Progress and Challenges. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP*, pages 2637–2652.

Satoru Uchida, Shohei Takada, and Yuki Arase. 2018. CEFR-based Lexical Simplification Dataset. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 3254–3258.

Rodrigo Wilkens, Leonardo Zilio, Silvio Ricardo Cordeiro, Felipe S.F. Paula, Carlos Ramisch, Marco A.P. Idiart, and Aline Villavicencio. 2017. LexSubNC: A Dataset of Lexical Substitution for Nominal Compounds. In *Proceedings of the International Conference on Computational Semantics (IWCS)*, volume Short papers.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in Current Text Simplification Research: New Data Can Help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Wei Xu, Chris Callison-Burch, and Courtney Nápoles. 2016. Optimizing Statistical Machine Translation for Text Simplification. In *Transactions of the Association for Computational Linguistics*, volume 4, pages 401–415.

Daiki Yanamoto, Tomoki Ikawa, Tomoyuki Kajiwara, Takashi Ninomiya, Satoru Uchida, and Yuki Arase. 2022. Controllable Text Simplification with Deep Reinforcement Learning. In *Proceedings of the Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (AACL) and the International Joint Conference on Natural Language Processing (IJCNLP)*, volume 2: Short Papers, pages 398–404.

Tatsuya Zetsu, Tomoyuki Kajiwara, and Yuki Arase. 2022. Lexically Constrained Decoding with Edit Operation Prediction for Controllable Text Simplification. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR)*, pages 147–153.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Yang Zhang, Hanlei Jin, Dan Meng, Jun Wang, and Jinghua Tan. 2026. A Comprehensive Survey

on Automatic Text Summarization with Exploration of LLM-based Methods. *Neurocomputing*, 663:131928.

Automatic Text Simplification for French Medical Documents with LLMs: The Role of Target Audience and Genre

Rémi Cardon[♦*], **A. Seza Doğruöz**[♥]
♦Computer Science and Engineering Department, UC3M, Spain
♥LT3, IDLab, Universiteit Gent, Belgium
rcardon@inf.uc3m.es, as.dogruoz@ugent.be

Abstract

Medical information is hard for non-specialists to understand, despite its importance for treatment success. Automatic text simplification (ATS) rewrites complex documents into simpler versions, with effectiveness measured through ATS evaluation metrics and readability metrics. A key challenge in ATS is calibrating simplification to match the reading abilities of specific target audiences, as different populations have different comprehension needs. Since socio-demographic factors such as education level and health literacy are known to correlate with reading abilities, we hypothesize that large language models (LLMs) may be able to adjust their simplification strategies when provided with descriptions of target audiences. In this study, we investigate how LLMs simplify French medical documents when prompted with socio-demographic characteristics of target patients. We compare this approach with prompts based on language proficiency levels (CEFR) to determine whether LLMs respond differently to explicit proficiency levels versus implicit audience descriptions. Our experiments with five LLMs on three types of French medical documents show that CEFR prompts produce greater readability variation (particularly for Llama-3.1-8B), while socio-demographic factors yield more homogeneous outputs. Text genre also considerably impacts LLM outputs for ATS.

1. Introduction

The ability to understand medical information is correlated with the chances of success for a medical treatment (Berkman et al., 2011). However, the growing availability of online medical information does not yield an improvement in understanding of medical content by the general public (Parker et al., 1999).

Automatic text simplification (ATS) is a line of research in natural language processing (NLP) that develops tools and resources to make written texts easier to read for target audiences (Saggion, 2017; Alva-Manchego et al., 2020), for example by shortening sentences or replacing technical terms. To assess whether ATS systems reduce comprehension difficulty, researchers rely on both ATS-specific evaluation metrics (e.g., comparing n-gram operations and measuring meaning preservation) and readability metrics that quantify textual characteristics as proxies for reading ease. However, sociodemographic characteristics (e.g., age, education level, health literacy) of the target audience are rarely taken into account in ATS research (Gooding, 2022).

In terms of implementation, ATS methods have recently been explored with large language models (LLMs) (Kew et al., 2023), which use prompts to simplify texts and increase their readability. ATS researchers have explored adding readability information to prompts (Barayan et al., 2025). To build upon this research, we investigate how LLMs simplify medical texts in French when the prompt includes information about the target audience. More precisely, we focus on the effect of target audience description on readability when simplifying medical texts in French.

Our hypothesis is that LLMs encode representations linking socio-demographic factors to language skills, particularly reading abilities. If validated, this would enable more targeted simplification strategies tailored to specific patient populations through audience-specific prompts. Our contributions are as follows: (i) We conduct, to our knowledge, the first study to investigate how socio-demographic descriptions of target audiences as patients influence the simplified medical text output of LLMs for French documents; (ii) we present, to our knowledge, the first study on document-level ATS for French medical texts.

*Research done while employed at UGent.

Proceedings of the Joint Workshop on Readability and Text Simplification (READIxTSAR) @ LREC 2026, pages 164–180
11 May 2026.

2. Related Work

2.1. Automatic Text Simplification

ATS has traditionally been performed at the sentence level (Chandrasekar et al., 1996; Alva-Manchego et al., 2020), taking a sentence as input and simplifying it to convey the same meaning in an easier-to-read form. For example, "*Medication inhibiting the peristalsis are counter-indicated in this situation*" can be simplified to "*In this case, do not take medication for stopping or decreasing the intestinal transit*" (Grabar and Saggion, 2022).

For English, efforts have extended beyond the sentence level to paragraphs (Devaraj et al., 2021; Lu et al., 2023) and full documents (Cripwell et al., 2023a; Mo and Hu, 2024; Nagai et al., 2024). However, medical ATS in French remains limited and has only been explored at the sentence level (Cardon and Grabar, 2020; Todirascu et al., 2022).

Regarding methodologies, the field has transitioned from rule-based systems to generative models over time, incorporating statistical approaches and other machine learning techniques (Cardon and Bibal, 2023).

ATS and Target Audience Most ATS work approaches simplification as a one-size-fits-all task (Gooding, 2022), despite the fact that different target audiences have different needs (Rennes et al., 2022). This monolithic approach is reflected in evaluation protocols. Automatic metrics require annotated data, either as references (Xu et al., 2016; Papineni et al., 2002) or as training data (Maddela et al., 2023; Cripwell et al., 2023b). Such corpora are costly to create, and available corpora are typically not associated with specific target audiences.

Regarding human evaluation of ATS, it is mostly carried out by researchers, their colleagues, or crowd workers, which is problematic in terms of representativeness of the target population (Doğruöz et al., 2023). A few studies have directly involved members of target audiences in evaluation (Alonzo et al., 2020, 2022), but this remains rare.

Readability-controlled ATS Readability-controlled ATS focuses on leveraging readability information to adjust LLM outputs to target readability levels. Common readability measures include (1) the Flesch-Kincaid Grade Level (Kincaid et al., 1975, FKGL), a traditional formula for English readability that uses the total number of words, sentences, and syllables to output a grade corresponding to estimated readability; and (2) the Common European Framework of Reference for Languages (Council of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division, 2001, CEFR), which assesses language proficiency for second language learners using six levels (A1, A2, B1, B2, C1, C2) with descriptions of expected skills at each level.

Recent work has incorporated FKGL or CEFR levels in prompts for simplification (Imperial and Tayyar Madabushi, 2023; Barayan et al., 2025), though only for English and sentence-level simplification. Moreover, these approaches consider readability as a textual property, leaving aside the socio-demographic characteristics of the reader.

Evaluation To evaluate ATS performance for document-level simplification, Sun et al. (2021) introduce D-SARI, an adaptation of SARI (Xu et al., 2016, System output Against References and Input), originally developed for sentence-level ATS. SARI compares token n-grams that are added, deleted, and kept when simplifying from input to reference text against the same operations from input to output texts. D-SARI follows the same principles with adjustments to penalize large discrepancies in document length and sentence repetitions.

Mo and Hu (2024) assess documents using readability features including traditional metrics (e.g., FKGL) and other aspects such as type-token ratio and syntactic complexity. ATS research commonly uses BERTScore (Zhang et al., 2020) to evaluate meaning preservation (Alva-Manchego et al., 2021). BERTScore is a metric that relies on BERT embeddings to measure semantic similarity between texts.

Regarding human evaluation, sentence simplification is typically judged in terms of grammaticality or fluency (*is the output grammatical?*), simplicity (*is the output simpler than the input?*), and meaning preservation or adequacy (*is the original meaning preserved?*)

using a 5-point Likert scale. This evaluation approach is also used for document-level simplification (Sun et al., 2021), though implementation and interpretation details are not yet stabilized (Stodden, 2021).

2.2. LLMs and Target Audiences

There is a lack of research on whether LLMs can effectively take socio-demographic factors into account. When addressing socio-demographic groups and LLMs, current NLP research mostly focuses on biases. For example, Navigli et al. (2023); Nozza et al. (2022); Kotek et al. (2023); Gallegos et al. (2024) identify bias through sentence completion tasks, prompting LLMs with the beginning of a group definition and studying the representation the model outputs. This research focuses on semantic content rather than linguistic form.

We found only one study focusing on the links between target groups and writing style (Malik et al., 2024). They incorporate socio-demographic factors (location, occupation, political affiliation, and age) in prompts and analyze the writing style of models when instructed to impersonate someone with those factors. In this section, we have reviewed studies on how LLMs behave when producing text *about* or *as* members of given social groups. We have found no study on how LLMs produce text *for* specific social groups, which is the focus of our question.

3. Methodology

In this section, we describe the data we use (Section 3.1), the models we selected for our experiments (Section 3.2), the socio-demographic factors we incorporate into our study (Section 3.3), and the metrics we use to analyze the outputs (Section 3.4).

3.1. Data

For our experiments, we use the CLEAR corpus (Grabar and Cardon, 2018), a freely available French medical corpus compiled for ATS research. It includes document pairs on the same topic targeting different audiences, with three different document types:

1. **Cochrane Summaries**: These are summaries written by the Cochrane Foundation[1] for medical practitioners and manually simplified into plain language summaries (PLS). Each summary addresses a specific medical research question. These summaries have also been used in biomedical ATS for English (Devaraj et al., 2021). The summaries are available in 20 different languages, including French.[2]

2. **Drug Information**: For every drug introduced to the French market, manufacturers must publicly release information[3] in two forms: one for medical practitioners (RCP, *résumé des caractéristiques du produit*, summary of product characteristics) and one for patients (the paper leaflet provided in each medicine box).

3. **Encyclopedia Articles**: Articles from the medical section of French Wikipedia[4] and corresponding articles from French Vikidia,[5] a collaborative online encyclopedia written for children aged 8 to 13.

For our experiments, we randomly sample 100 document pairs (complex and simple versions) from each sub-corpus. Examples for each dataset are available in Appendix B.

3.2. Models

We experiment with five open-source LLMs of comparable size (7-9B parameters), which all support French and have instruction-following capabilities: `Llama-3.1-8B-Instruct`[6] (henceforth `Llama`), `DeepSeek-R1-Distill-Llama-8B`[7] (henceforth `LlamaDS`), a version of Llama

[1] https://www.cochrane.org/

[2] https://cochranesupport.wiley.com/s/article/languages-supported-in-cochrane-library

[3] https://base-donnees-publique.medicaments.gouv.fr/

[4] https://fr.wikipedia.org/wiki/Portail:Médecine

[5] https://fr.vikidia.org/wiki/Portail:Médecine

[6] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[7] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-8B

3.1 fine-tuned on DeepSeek's outputs, `Qwen2.5-7B-Instruct`[8] (henceforth `Qwen`), `DeepSeek-R1-Distill-Qwen-7B`[9] (henceforth `QwenDS`), a version of Qwen 2.5 fine-tuned on DeepSeek's outputs, and `BioMistral-7B`[10] (Labrak et al., 2024), a biomedical model.

These models were selected based on several criteria: (1) comparable size to ensure fair comparison, (2) availability and reproducibility as open-source models, (3) documented performance on multilingual or French tasks, and (4) for BioMistral, specialization in the biomedical domain. For all experiments, we set the temperature to 0.01, ensuring that the same prompt yields consistent outputs. On average, each model took approximately 16 hours per corpus (approximately 48 hours total across all three corpora) to process 100 documents with the 13 different prompts (6 CEFR levels, 5 IPS-based factors, and 2 health literacy conditions) on an NVIDIA GeForce RTX 4090.

3.3. Socio-demographic Factors

The socio-demographic factors in our study are selected based on their documented relationship with the ability to access and process medical information, which we use as an approximation of reading abilities. We focus on the level of education and health literacy.

Level of Education To indicate the level of education (for the reader), we rely on data released by the French government. For each elementary school (*école primaire*), middle school (*collège*), and high school (*lycée*) in France, the French Ministry of Education publishes an indicator called IPS (*indice de position sociale*, social position index), which measures the social status of students (Dauphant et al., 2023). IPS is calculated using PCS (*profession et catégorie sociale*[11]), which maps 40 parental occupations to numeric codes. IPS is calculated for each student based on the PCS value assigned to their parent(s).[12] IPS values range from 45 (a student with an unqualified working mother and a student father) to 185 (a student with an engineer mother and a professor/scientist father). PCS codes have different weights depending on whether the parent is a mother or a father (Rocher, 2016). For example, a student with a single mother who is a teacher has an IPS of 154, while a student with a single father with the same occupation has an IPS of 146.[13] The publicly released values are average IPS per school type.

IPS is used to inform policy decisions regarding education in France. Schools with low average IPS receive additional public funding (e.g., higher teacher-to-student ratios, educational assistants). IPS values correlate with diploma success rates (e.g., baccalauréat) and grade progression, strengthening IPS as a relevant indicator for approximating students' abilities to access and process information. IPS values also correlate with high school type ("secteur" in French): public schools (free registration, only public funding) or private schools (paid registration, partial public funding).

Health Literacy Health literacy (HL) in France refers to patients' ability to access, understand, evaluate, and use information needed for healthcare (Rey et al., 2023). According to this study, 10% of the French population has low HL, rising to 33% among people with self-declared 'poor' or 'very poor' health. Additionally, HL increases with education level (i.e., the level of official diploma obtained). This evidence reinforces the relationship between reading abilities and socio-demographic factors of patient populations. The connection between HL and reading abilities is particularly relevant for ATS: patients with lower HL struggle not only with medical-specific content but also with the linguistic complexity of health documents. Self-reported health status serves as a proxy for HL in our study because the Rey et al. study demonstrates a strong correlation between poor health and low HL (33% vs. 10% in the general population).

[8] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

[9] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-7B

[10] https://huggingface.co/BioMistral/BioMistral-7B

[11] occupation and social category

[12] https://www.education.gouv.fr/l-indice-de-position-sociale-ips-357755

[13] https://www.education.gouv.fr/media/158757/download

Factors Used in Prompts Building upon these studies on French education and health, we include the following social factors in our prompts:

- **High school students** (based on IPS studies):
  - High school type: *lycée public / privé* (public, private)
  - Household income level: *milieu défavorisé / moyen / favorisé* (low, average, high)
- **Adults** (based on HL studies):
  - Health condition: *mauvais / bon état de santé* (poor, good)

In addition, to facilitate comparison with recent ATS approaches (see Section 2), we include the six CEFR levels (A1, A2, B1, B2, C1, C2), a scale of language proficiency levels.

3.4. Analysis Metrics

To analyze the readability of texts in our corpus and the models' outputs, we use the following metrics.

Reading Ease Level (REL) We compute the Reading Ease Level (Kandel and Moles, 1958, REL), an adaptation of Flesch Reading Ease (Flesch, 1948) for French. A higher score indicates a more readable text. The formula is:

$$207 - 1.015 \times \frac{\text{Number of words}}{\text{Number of sentences}} - 73.6 \times \frac{\text{Number of syllables}}{\text{Number of words}} \quad (1)$$

Automatic Evaluation Metrics We use two automatic metrics: D-SARI[14], which measures simplification quality by comparing n-gram operations between input, output, and reference texts (Sun et al., 2021), and BERTScore[15] (using the `bert-base-multilingual-cased` model), which measures meaning preservation through semantic similarity (Zhang et al., 2020).

[14]Implementation: https://github.com/RLSNLP/Document-level-text-simplification

[15]Implementation: https://github.com/Tiiiger/bert_score

4. Results

We begin by examining the original CLEAR corpus texts in terms of readability characteristics and automatic metrics. We then analyze the outputs of our selected models, first providing an overview of their performance, then examining the impact of text genre and prompt factors on simplification outcomes.

4.1. Original Texts

We first examine the simplification strategies present in the human-written simplifications of the CLEAR corpus. Figure 1 shows the distribution of REL score differences between complex and simple text pairs, broken down by dataset.[16] Higher values indicate greater simplification according to REL.

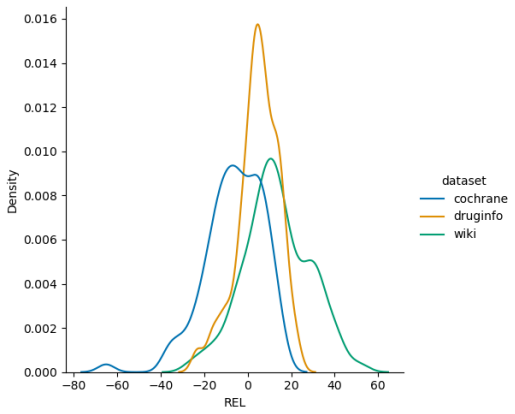


Figure 1: Difference in REL values between complex and simple versions of CLEAR subcorpora. Higher values indicate greater simplicity gain.

We observe that simplification strategies differ substantially between datasets. The Drug Information (Druginfo) dataset shows the least variation, with differences centered near zero. The encyclopedia dataset (Wiki) shows an increase in REL values (indicating successful simplification), whereas the Cochrane reviews show a decrease (indicating that the simplified versions have lower REL scores than the original texts). This suggests that different text

[16]All figures in this paper use `seaborn`'s `colorblind` color palette (https://seaborn.pydata.org/generated/seaborn.color_palette.html).

genres employ different simplification strategies, and that REL alone may not capture all aspects of simplification, particularly for specialized medical literature.

4.2. Model Simplifications

4.2.1. Overview and Data Cleaning

We plot the REL values by model across all datasets and prompts (Figure 2). While the lowest REL value in human-written texts is 4.95 (a Cochrane review on antibiotics for outpatient treatment), the models sometimes output texts with negative REL values. There are 160 such texts across the 19,500 generated texts: 97 produced by `BioMistral`, 51 by `QwenDS`, and 12 by `Llama`. These anomalous outputs are spread equally across all prompts.

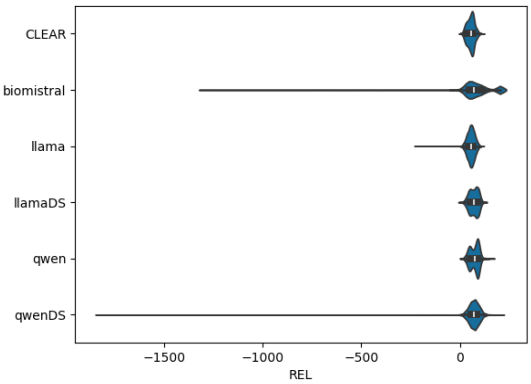


Figure 2: REL values by model and for the original simple documents (CLEAR), across all datasets and prompts.

The lowest REL value (-1,826.45) is produced by `QwenDS` on a simplification of the “Ectopia" Wikipedia article with the *poor health* prompt. The output ends with 5,496 consecutive characters composed of repetitions of non-French character sequences. Manual inspection reveals that texts with REL values below 4 contain similar anomalies (repetitions, formatting errors, or truncated outputs).

On the other hand, the highest REL value for human-written texts is 115.41 (a Vikidia article on the urinary meatus). The maximum value for LLM-generated texts (205.96) is obtained by `BioMistral` on 367 outputs, all consisting of a single number (digit 1 followed by zeros). Among the 847 texts with REL above 130, 799 are produced by `BioMistral` (20.49%), 28 by `QwenDS` (0.72%), and 20 by `Qwen` (0.51%). Notably, the `Qwen` models' high-REL outputs are partially or entirely in Chinese despite being prompted in French.

Based on these observations, we filter out texts with REL 4 or REL 130. The percentage of texts retained after filtering: `BioMistral` 76.84%, `Llama` 99.67%, `LlamaDS` 99.95%, `Qwen` 99.49%, and `QwenDS` 97.97%. As we identified almost 25% of anomalous texts from `BioMistral`, we exclude this model from subsequent analyses. The lower bound of 4 corresponds to the minimum REL observed in human-written texts; the upper bound of 130 was chosen with a conservative margin above the maximum human REL (115.41). Future work should investigate more principled threshold-selection methods.

4.2.2. Impact of Text Genre

Figure 3 shows the distribution of REL values after filtering, broken down by dataset. For all models, the range of REL values is wider than for the original simple documents. However, the genre of the original text clearly impacts the output. All models produce texts with lower REL values for the Cochrane dataset than for the other two. `Llama` shows the greatest proximity to human-written texts, while the REL distributions for the three other models appear more similar to each other.

Table 1 shows D-SARI and BERTScore by model and dataset. D-SARI scores align with those reported in recent document-level ATS work on English datasets (Fang et al., 2025; Bahrainian et al., 2024). Given that we operate in a zero-shot setting and that simplification strategies vary considerably across the three corpora (as shown in Figure 1), we interpret these scores as evidence that model outputs warrant further investigation despite the absence of corpus-specific fine-tuning.

In terms of meaning preservation (BERTScore), `LlamaDS` produces texts closest to the original across all datasets, even exceeding human simplifications for Cochrane and Wiki. D-SARI is not reported for human simplifications (CLEAR) because the simple documents in CLEAR serve as the reference texts in this metric, making

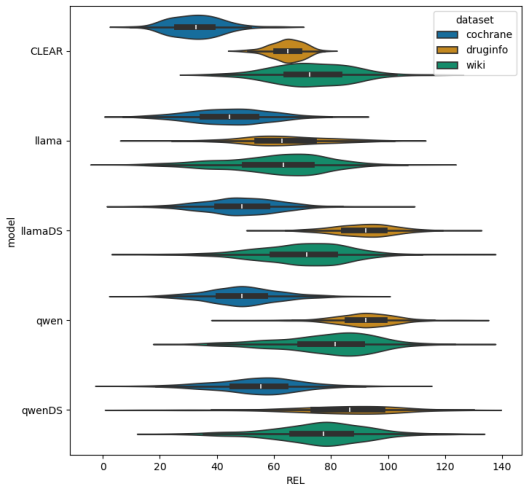


Figure 3: REL values between 4 and 130, by model and for the original simple documents (CLEAR), across all datasets and prompts.

a comparison of reference against itself undefined. Conversely, `Qwen` obtains the lowest BERTScore values. The Druginfo dataset proves most challenging for all models, though notably, human simplifications for this dataset retain meaning better than for the other two.

4.2.3. Impact of Factors in the Prompt

To examine the impact of prompt factors on outputs, we analyze REL and BERTScore values. Figure 4 (Appendix A) shows these values for each model across all datasets, with factors grouped by type: high school students (based on IPS studies), adults (based on health literacy studies), and language proficiency (CEFR levels).

CEFR levels have the most substantial influence on outputs. `Llama` shows the most pronounced progression from A1 to C2 for both REL and BERTScore, with A1 prompts producing the most readable texts (highest REL) and C2 prompts producing the least readable texts (lowest REL), consistent with the scale's intended progression. `QwenDS` is the only model that does not appear to have encoded the CEFR scale, showing no systematic variation across levels.

All models show minimal sensitivity to factors related to health condition (good vs. poor health). Similarly, income-related prompts (IPS) show little effect, though `Llama` produces texts with notably higher REL values for the *low income* factor compared to *average* and *high income*.

	D-SARI	BERTScore
CLEAR (human)		
Cochrane	-	73.75
Druginfo	-	76.60
Wiki	-	69.70
Llama		
Cochrane	**35.57**	73.14
Druginfo	29.51	60.84
Wiki	32.68	72.30
LlamaDS		
Cochrane	31.78	**79.83**
Druginfo	**31.01**	**65.14**
Wiki	30.49	**74.79**
Qwen		
Cochrane	34.07	68.40
Druginfo	28.51	60.91
Wiki	**33.92**	69.43
QwenDS		
Cochrane	30.74	77.38
Druginfo	29.91	61.73
Wiki	28.61	70.91

Table 1: D-SARI and BERTScore values for each model by dataset, across all prompts. Bold values indicate highest scores for each dataset.

5. Discussion

Our results highlight substantial variation in how models respond to prompt factors. `Llama` shows the greatest sensitivity, with outputs complying with traditional readability definitions when prompted with CEFR levels. Other models behave differently: their CEFR outputs show lower REL variation, and the DeepSeek-distilled models (`LlamaDS` and `QwenDS`) do not strictly follow the scale (e.g., *A1* produces lower REL values than *A2*).

Socio-demographic factors impact readability differently than CEFR levels. Models produce limited variation across factors like high school type, income level, health condition. However, there is one notable exception. The

low income prompt yields the highest REL values among socio-demographic factors for `Llama`. It is the only factor that correlates with both lower education and low health literacy (Dauphant et al., 2023; Rey et al., 2023). This finding suggests that the model has captured this relationship.

Conversely, *poor health* does not trigger highly readable outputs, instead producing the second-lowest REL values after *good health*. This suggests models' representations of socio-demographic factors and reading abilities may be incomplete or inconsistent.

5.1. Impact of Text Genre

Our results demonstrate that text genre has a considerable impact on model outputs. All models produce texts with lower REL values (indicating less readability according to this metric) for the Cochrane dataset compared to the other two datasets. This finding is particularly significant given that the human-written simplifications in the CLEAR corpus also show genre-specific patterns (Figure 1).

This genre effect carries important implications for ATS system design, suggesting that simplification strategies should adapt to both target audiences and source text genres. Even when targeting the same audience, different document types—medical literature reviews (Cochrane), drug information leaflets (Druginfo), and encyclopedia articles (Wiki)—may benefit from distinct simplification approaches. Future work should investigate whether explicitly incorporating genre information in prompts enhances simplification quality.

5.2. Model Comparison and Fine-tuning Effects

The differences between `Llama` and `Qwen`, and their respective DeepSeek-distilled counterparts (`LlamaDS` and `QwenDS`), highlight the significant impact of fine-tuning strategies on model capabilities for ATS. `QwenDS` shows no systematic variation when prompted with different factors, suggesting that the distillation process may have reduced the model's sensitivity to prompt nuances. This finding underscores the importance of careful evaluation when adopting fine-tuned or distilled models for tasks requiring prompt-based control.

`LlamaDS` achieves the highest BERTScore values across all datasets, indicating superior meaning preservation compared to other models and even surpassing human simplifications for some datasets. However, this comes at the cost of reduced readability variation in response to prompt factors. This trade-off between meaning preservation and controllable simplification is an important consideration for practical ATS systems.

6. Conclusion

We investigated how text genre and socio-demographic factors in prompts affect outputs when prompting LLMs for document-level ATS of French medical texts. `Llama-3.1-8B` shows the greatest sensitivity to prompt factors, while other models show less variation or inconsistent patterns.

For `Llama`, CEFR levels in prompts produce outputs that comply with traditional definitions of text complexity, following the expected progression from A1 (most readable) to C2 (least readable). Socio-demographic factors trigger different patterns, with the *low income* factor producing notably more readable outputs, consistent with documented correlations between socioeconomic status and reading abilities. Other socio-demographic factors (health condition, school type) show minimal effect.

We also demonstrate that text genre has a considerable impact on model outputs, with all models producing different readability characteristics for Cochrane reviews, drug information, and encyclopedia articles. This finding suggests that effective ATS systems should account for both target audience characteristics and source text genre.

Our findings represent a first step toward systematically incorporating user descriptions in LLM prompts for ATS. Future work should investigate additional socio-demographic factors, test interactions between multiple factors, diversify the array of models evaluated, and most importantly conduct validation studies with members of target populations to assess whether the observed readability variations translate to improved comprehension and usability.

7. Limitations

As our work explores novel research questions, the scope of our experiments is necessarily limited. We selected five models of comparable size to enable fair comparison, but focused detailed analysis on four models after excluding `BioMistral` due to a high rate of anomalous outputs. The rapid evolution of LLMs means that our findings may not generalize to newer or larger models.

Our study shares limitations common to ATS research. Little is known about the concrete effectiveness of different simplification strategies for specific target audiences. While we found readability variation depending on sociodemographic factors in prompts, we cannot confirm whether these variations are actually beneficial for the corresponding populations. This limitation is particularly significant given the unexpected behavior for some factors (e.g., *poor health* producing less readable outputs than *good health*).

We did not conduct human evaluation with members of target populations. Automatic metrics (REL, D-SARI, BERTScore) provide useful quantitative measures but do not capture all aspects of text comprehension and usability. For instance, REL is based primarily on sentence and word length, which may not fully reflect the complexity of medical terminology or conceptual difficulty. Conducting studies with actual target audiences represents important future work that will determine whether our quantified readability variations correspond to meaningful improvements in understanding.

Additionally, our analysis focused on readability metrics and did not examine other aspects of simplification quality, such as the preservation of medical accuracy, the appropriateness of added elaborations, or the completeness of information. The exclusion of linguistic criteria analysis (e.g., passive voice, sentence complexity patterns) limits our understanding of specific simplification strategies employed by different models.

8. Plain Language Summary

Understanding medical information can be difficult for people who are not medical experts. This is a problem because patients need to understand their health information to make good decisions about their care. One way to help is by using computer programs that rewrite complicated medical texts into simpler language. These programs are called automatic text simplification (ATS) systems.

In our study, we wanted to see if artificial intelligence (AI) models could make French medical documents easier to read for different groups of people. We tested whether these AI models could change the way they simplify text based on who the reader is. For example, we gave the AI information about the reader's education level or health background, or we told it what level of French the reader understands.

We used five different AI models and three types of French medical documents. We found that when we told the AI the reader's language skill level, the simplified texts varied more in how easy they were to read. When we described the reader's background (like education or health), the AI's simplified texts were more similar to each other. The type of medical document also made a difference in how the AI rewrote the text.

In summary, our research shows that AI can help make medical information easier to understand, and that the way we describe the intended reader affects how the AI rewrites the text. This could help create better health information for different groups of people.

9. Bibliographical References

Oliver Alonzo, Sooyeon Lee, Mounica Maddela, Wei Xu, and Matt Huenerfauth. 2022. A dataset of word-complexity judgements from deaf and hard-of-hearing adults for text simplification. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 119–124, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Oliver Alonzo, Matthew Seita, Abraham Glasser, and Matt Huenerfauth. 2020. Automatic text simplification tools for deaf and hard of hearing adults: Benefits of lexical simplification and providing users with au-

tonomy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, NY, USA. Association for Computing Machinery.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020. Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, 46(1):135–187.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. The (un)suitability of automatic evaluation metrics for text simplification. *Computational Linguistics*, 47(4):861–889.

Seyed Ali Bahrainian, Jonathan Dou, and Carsten Eickhoff. 2024. Text simplification via adaptive teaching. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6574–6584, Bangkok, Thailand. Association for Computational Linguistics.

Abdullah Barayan, Jose Camacho-Collados, and Fernando Alva-Manchego. 2025. Analysing zero-shot readability-controlled sentence simplification. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6762–6781, Abu Dhabi, UAE. Association for Computational Linguistics.

Nancy D Berkman, Stacey L Sheridan, Katrina E Donahue, David J Halpern, and Karen Crotty. 2011. Low health literacy and health outcomes: an updated systematic review. *Annals of internal medicine*, 155(2):97–107.

Rémi Cardon and Adrien Bibal. 2023. On operations in automatic text simplification. In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 116–130, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Rémi Cardon and Natalia Grabar. 2020. French biomedical text simplification: When small and precise helps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 710–716.

R. Chandrasekar, Christine Doran, and B. Srinivas. 1996. Motivations and methods for text simplification. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*.

Council of Europe. Council for Cultural Cooperation. Education Committee. Modern Languages Division. 2001. *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2023a. Document-level planning for text simplification. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 993–1006, Dubrovnik, Croatia. Association for Computational Linguistics.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2023b. Simplicity level estimate (SLE): A learned reference-less metric for sentence simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12053–12059, Singapore. Association for Computational Linguistics.

Fannie Dauphant, Franck Evain, Marine Guillerm, Catherine Simon, and Thierry Rocher. 2023. L'indice de position sociale (ips): Un outil statistique pour décrire les inégalités sociales entre établissements. *Note d'information de la DEPP*, pages pp–1.

Ashwin Devaraj, Iain Marshall, Byron Wallace, and Junyi Jessy Li. 2021. Paragraph-level simplification of medical texts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4972–4984, Online. Association for Computational Linguistics.

A. Seza Doğruöz, Sunayana Sitaram, and Zheng Xin Yong. 2023. Representativeness as a forgotten lesson for multilingual and code-switched data collection and preparation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5751–5767, Singapore. Association for Computational Linguistics.

Dengzhao Fang, Jipeng Qiang, Xiaoye Ouyang, Yi Zhu, Yunhao Yuan, and Yun Li. 2025. Collaborative document simplification using multi-agent systems. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 897–912, Abu Dhabi, UAE. Association for Computational Linguistics.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, pages 1–79.

Sian Gooding. 2022. On the ethical considerations of text simplification. In *Ninth Workshop on Speech and Language Processing for Assistive Technologies (SLPAT-2022)*, pages 50–57, Dublin, Ireland. Association for Computational Linguistics.

Natalia Grabar and Rémi Cardon. 2018. CLEAR – simple corpus for medical French. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, pages 3–9, Tilburg, the Netherlands. Association for Computational Linguistics.

Natalia Grabar and Horacio Saggion. 2022. Evaluation of automatic text simplification: Where are we now, where should we go from here. In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, pages 453–463, Avignon, France. ATALA.

Joseph Marvin Imperial and Harish Tayyar Madabushi. 2023. Flesch or fumble? evaluating readability standard alignment of instruction-tuned language models. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 205–223, Singapore. Association for Computational Linguistics.

Liliane Kandel and Abraham Moles. 1958. Application de l'indice de flesch à la langue française. *Cahiers Etudes de Radio-Télévision*, 19(1958):253–274.

Tannon Kew, Alison Chi, Laura Vásquez-Rodríguez, Sweta Agrawal, Dennis Aumiller, Fernando Alva-Manchego, and Matthew Shardlow. 2023. BLESS: Benchmarking large language models on sentence simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13291–13309, Singapore. Association for Computational Linguistics.

J Peter Kincaid, RP Fishburne, RL Rogers, and BS Chissom. 1975. Derivation of new readability formulas (automated reliability index, fog count and flesch reading ease formula) for navy enlisted personnel (research branch report 8-75). memphis, tn: Naval air station; 1975. *Naval Technical Training, US Naval Air Station: Millington, TN*.

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference*, pages 12–24.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains.

Junru Lu, Jiazheng Li, Byron Wallace, Yulan He, and Gabriele Pergola. 2023. NapSS: Paragraph-level medical text simplification via narrative prompting and sentence-matching summarization. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1079–1091, Dubrovnik, Croatia. Association for Computational Linguistics.

Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2023. LENS: A learnable evaluation metric for text simplification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16383–16408, Toronto, Canada. Association for Computational Linguistics.

Manuj Malik, Jing Jiang, and Kian Ming A. Chai. 2024. An empirical analysis of the writing styles of persona-assigned LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19369–19388, Miami, Florida, USA. Association for Computational Linguistics.

Kaijie Mo and Renfen Hu. 2024. ExpertEase: A multi-agent framework for grade-specific document simplification with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9080–9099, Miami, Florida, USA. Association for Computational Linguistics.

Yoshinari Nagai, Teruaki Oka, and Mamoru Komachi. 2024. A document-level text simplification dataset for Japanese. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 459–476, Torino, Italia. ELRA and ICCL.

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: origins, inventory, and discussion. *ACM Journal of Data and Information Quality*, 15(2):1–21.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2022. Pipelines for social bias testing of large language models. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 68–74, virtual+Dublin. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ruth M Parker, Mark V Williams, Barry D Weiss, David W Baker, Terry C Davis, Cecilia C Doak, Leonard G Doak, Karen Hein, Cathy D Meade, Joann Nurss, et al. 1999. Health literacy-report of the council on scientific affairs. *Jama-Journal of the American Medical Association*, 281(6):552–557.

Evelina Rennes, Marina Santini, and Arne Jonsson. 2022. The Swedish simplification toolkit: – designed with target audiences in mind. In *Proceedings of the 2nd Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI) within the 13th Language Resources and Evaluation Conference*, pages 31–38, Marseille, France. European Language Resources Association.

Sylvie Rey, Aude Leduc, Xavier Debussche, Laurent Rigal, V Ringa, V Costemalle, et al. 2023. Une personne sur dix éprouve des difficultés de compréhension de l'information médicale. *Études et résultats*, 1269(mai):8.

Thierry Rocher. 2016. Construction d'un indice de position sociale des élèves. *Éducation & formations*, (90):5–27.

H. Saggion. 2017. *Automatic Text Simplification*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Regina Stodden. 2021. When the scale is unclear–analysis of the interpretation of rating scales in human evaluation of text simplification. *Proceedings of the 1st Workshop on Current Trends in Text Simplification (CTTS 2021, co-located with SEPLN 2021)*.

Renliang Sun, Hanqi Jin, and Xiaojun Wan. 2021. Document-level text simplification: Dataset, criteria and baseline. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7997–8013, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

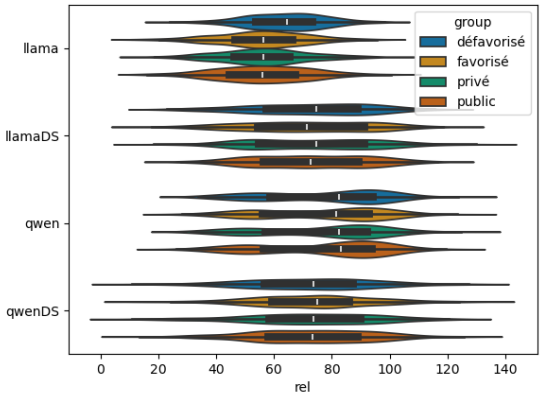
Amalia Todirascu, Rodrigo Wilkens, Eva Rolin, Thomas François, Delphine Bernhard, and Núria Gala. 2022. HECTOR: A hybrid TExt SimplifiCation TOol for raw texts in French. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4620–4630, Marseille,

France. European Language Resources Association.

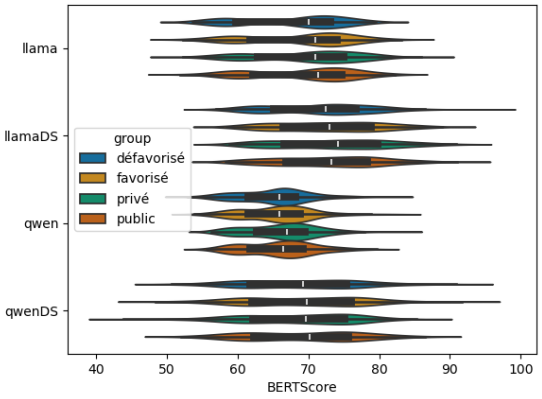
Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

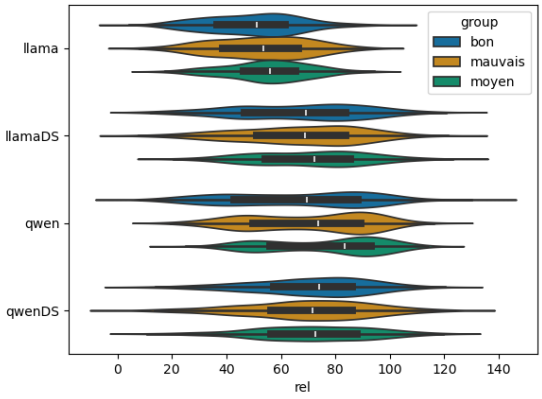
A. Results by Factor Group



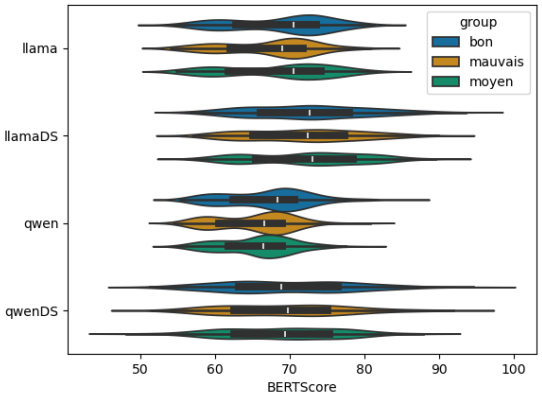
(a) IPS REL



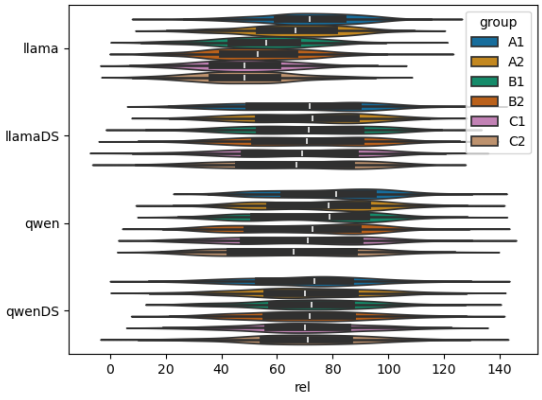
(b) IPS BERTScore



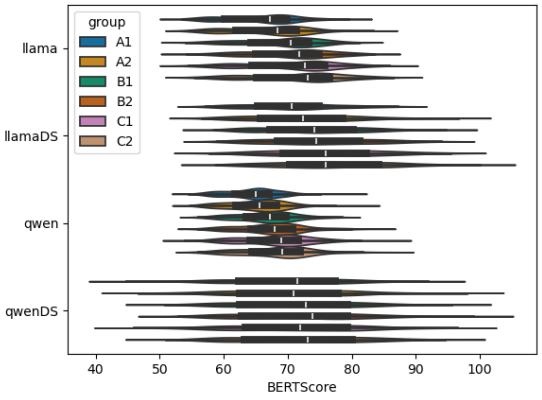
(c) HL REL



(d) HL BERTScore



(e) CEFR REL



(f) CEFR BERTScore

Figure 4: REL and BERTScore values for the different factors used in the prompts, by model.

B. Examples from CLEAR Corpus

Examples from each dataset type are provided in Tables 2, 3, and 4.

Cochrane review	
Literie en plume versus literie synthétique pour les patients asthmatiques. Contexte. Deux récentes études épidémiologiques rapportaient que les épisodes de respiration sifflante étaient plus fréquents chez les enfants utilisant des oreillers synthétiques que chez ceux utilisant des oreillers en plume. Objectifs. Évaluer l'efficacité de la literie en plume pour contrôler les symptômes de l'asthme. Stratégie de recherche documentaire. Le registre spécialisé du groupe Cochrane sur les voies respiratoires a été consulté en utilisant des termes prédéfinis. Les recherches étaient à jour en février 2009. Critères de sélection. Seuls les essais randomisés et les essais cliniques comparatifs étaient éligibles. Recueil et analyse des données. Aucun essai ne remplissait les critères d'inclusion dans la revue. Résultats principaux. La consultation de la littérature électronique a permis d'identifier 15 études en vue de l'examen de l'intégralité des articles. Après examen, aucune de ces études ne remplissait les critères d'inclusion dans la revue. Conclusions des auteurs. Bien que de récentes études épidémiologiques suggèrent que la literie en plume est associée à une respiration sifflante moins fréquente qu'avec des fibres synthétiques, les preuves actuellement disponibles sont insuffisantes pour évaluer les bénéfices cliniques de la literie en plume dans la prise en charge de l'asthme.	Feather versus synthetic bedding for patients with asthma. Background. Two recent epidemiological studies reported that wheezing episodes were more common in children using synthetic pillows than in those using feather pillows. Objectives. To assess the effectiveness of feather bedding in controlling asthma symptoms. Search strategy. The Cochrane Airways Group Specialised Register was searched using predefined terms. The searches were current to February 2009. Selection criteria. Only randomised trials and controlled clinical trials were eligible. Data collection and analysis. No trials met the inclusion criteria for the review. Main results. Electronic literature search identified 15 studies for full-text review. After review, none of these studies met the inclusion criteria for the review. Authors' conclusions. Although recent epidemiological studies suggest that feather bedding is associated with less frequent wheezing than synthetic fibres, the evidence currently available is insufficient to assess the clinical benefits of feather bedding in the management of asthma.
PLS	
Literie en plume versus literie synthétique pour les patients asthmatiques. Un allergène est une substance qui déclenche une réaction allergique chez les personnes qui y sont sensibles. Les acariens sont l'un des principaux allergènes de l'asthme. On pense que les oreillers et la literie contenant des fibres artificielles (fabriquées par l'homme) sont moins susceptibles d'accumuler des allergènes que les oreillers et les édredons en plume. Néanmoins, certaines preuves indiquent que la literie en plume pourrait, au contraire, être moins susceptible de causer de l'asthme. Cette revue n'a identifié aucun essai comparant les plumes aux fibres synthétiques et des recherches sont nécessaires afin d'établir le type de literie le mieux adapté aux patients asthmatiques.	Feather versus synthetic bedding for asthma patients. An allergen is a substance that triggers an allergic reaction in people who are sensitive to it. Dust mites are one of the main allergens in asthma. Pillows and bedding containing artificial (man-made) fibres are thought to be less likely to accumulate allergens than feather pillows and duvets. However, there is some evidence that feather bedding may be less likely to cause asthma. This review did not identify any trials comparing feathers with synthetic fibres and research is needed to establish which type of bedding is best for asthma patients.

Table 2: Example of a Cochrane review and its corresponding PLS, with English translations (done with Google Translate).

RCP	
4.3. Contre-indications Ce médicament est contre-indiqué en cas d'hypersensibilité à l'un des constituants. 4.4. Mises en garde spéciales et précautions d'emploi Le traitement par cet élément minéral trace ne dispense pas d'un traitement spécifique éventuel. Ce médicament contient du lactose. Son utilisation est déconseillée chez les patients présentant une intolérance au galactose, un déficit en lactase de Lapp ou un syndrome de malabsorption du glucose ou du galactose (maladies héréditaires rares). 4.5. Interactions avec d'autres médicaments et autres formes d'interactions Les données disponibles à ce jour ne laissent pas supposer l'existence d'interactions cliniquement significatives. 4.6. Grossesse et allaitement En l'absence de données expérimentales et cliniques et par mesure de précaution, l'utilisation de ce médicament est à éviter pendant la grossesse et l'allaitement.	4.3. Contraindications This medication is contraindicated in cases of hypersensitivity to any of the ingredients. 4.4. Special warnings and precautions for use Treatment with this trace mineral does not replace the need for specific treatment. This medication contains lactose. Its use is not recommended in patients with galactose intolerance, the Lapp lactase deficiency, or glucose-galactose malabsorption (rare hereditary diseases). 4.5. Interactions with other medicinal products and other forms of interaction The data available to date do not suggest the existence of clinically significant interactions. 4.6. Pregnancy and breastfeeding In the absence of experimental and clinical data, and as a precautionary measure, the use of this medication should be avoided during pregnancy and breastfeeding.
Leaflet	
Contre-indications Ne prenez jamais OLIGOSTIM COBALT, comprimé dans le cas suivant: · antécédent d'allergie à l'un des constituants. EN CAS DE DOUTE, IL EST INDISPENSABLE DE DEMANDER L'AVIS DE VOTRE MEDECIN OU DE VOTRE PHARMACIEN. Précautions d'emploi ; mises en garde spéciales Faites attention avec OLIGOSTIM COBALT, comprimé: Mises en garde spéciales Le traitement par cet élément minéral trace ne dispense pas d'un traitement spécifique éventuel. L'utilisation de ce médicament est déconseillée chez les patients présentant une intolérance au galactose, un déficit en lactase de Lapp ou un syndrome de malabsorption du glucose ou du galactose (maladies héréditaires rares). Précautions d'emploi EN CAS DE DOUTE NE PAS HESITER A DEMANDER L'AVIS DE VOTRE MEDECIN OU DE VOTRE PHARMACIEN. Interactions avec d'autres médicaments Prise ou utilisation d'autres médicaments: Si vous prenez ou avez pris récemment un autre médicament, y compris un médicament obtenu sans ordonnance, parlez-en à votre médecin ou à votre pharmacien.	Contraindications Never take OLIGOSTIM COBALT tablets in the following cases: · a history of allergy to any of the ingredients. IF IN DOUBT, IT IS ESSENTIAL TO ASK YOUR DOCTOR OR PHARMACIST FOR ADVICE. Precautions for use; special warnings Take special care with OLIGOSTIM COBALT tablets: Special warnings Treatment with this trace mineral does not replace the need for specific treatment. The use of this medicine is not recommended for patients with galactose intolerance, the Lapp lactase deficiency, or glucose-galactose malabsorption (rare hereditary diseases). Precautions for use IF IN DOUBT, DO NOT HESITATE TO ASK YOUR DOCTOR OR PHARMACIST FOR ADVICE. Interactions with other medications Taking or using other medications: If you are taking or have recently taken any other medication, including medication obtained without a prescription, talk to your doctor or pharmacist.

Table 3: Extracts of an RCP and its corresponding leaflet, with English translations (done with Google Translate).

Wikipedia article	
Les maladies cardio-vasculaires (ou maladies cardiovasculaires) sont les maladies qui concernent le cœur et la circulation sanguine. Dans les pays occidentaux, l'expression la plus courante est la maladie coronarienne, responsable de l'angine de poitrine ou encore des infarctus. Ces maladies touchent plus certaines catégories de population (ouvriers, personnes exposées à certaines pollutions, victimes d'obésité, etc.) et leur prévalence régionale est marquée (par exemple en France, à la fin du XXe siècle dans le Nord-Pas-de-Calais et en Alsace, deux régions nettement plus touchées que les autres régions et la moyenne nationale, comme pour plusieurs types de cancers)[1]. Elles comptent souvent parmi les facteurs qui diminuent le plus l'espérance de vie d'une population et semblent être un facteur de risque de dépression (chez les jeunes filles au moins[2]).	Cardiovascular diseases (or cardiovascular diseases) are diseases that affect the heart and blood circulation. In Western countries, the most common term is coronary artery disease, responsible for angina or heart attacks. These diseases affect certain population groups more (workers, people exposed to certain types of pollution, victims of obesity, etc.), and their regional prevalence is marked (for example, in France, at the end of the 20th century in Nord-Pas-de-Calais and Alsace, two regions significantly more affected than other regions and the national average, as is the case for several types of cancer)[1]. They are often among the factors that most reduce a population's life expectancy and appear to be a risk factor for depression (at least among young women[2]).
Vikidia article	
Une maladie cardio-vasculaire (ou cardiovasculaire) est une maladie qui touche le fonctionnement du cœur et de la circulation sanguine. Selon l'OMS, les maladies cardio-vasculaires représentent 29 % de la mortalité totale, ce qui en fait la première cause de mortalité dans le monde1. Les principales causes des maladies cardio-vasculaires sont le tabagisme, une mauvaise alimentation pas assez variée, et un manque d'activité sportive1. Faire suffisamment d'exercice chaque jour statistiquement divise le risque d'AVC et d'accident cardiaque par deux à tout âge.	Cardiovascular disease (or cardiovascular disease) is a condition that affects the functioning of the heart and blood circulation. According to the WHO, cardiovascular disease accounts for 29% of all deaths, making it the leading cause of death worldwide. The main causes of cardiovascular disease are smoking, an unhealthy and insufficiently varied diet, and a lack of physical activity. Getting enough exercise every day statistically halves the risk of stroke and cardiac arrest at any age.

Table 4: Extract of a Wikipedia article and its corresponding Vikidia article, with English translations (done with Google Translate).

A Learner-Oriented Annotated Resource of French Multiword Expressions for Text Adaptation in Foreign Language Reading

Anna Kalinina[1], Thomas François[2], Hélène Vassiliadou[1], Amalia Todirascu[1]

[1] University of Strasbourg, UR 1339/LiLPa & ITI LiRiC, [2] UCLouvain, CENTAL

[1] 22 rue René Descartes, BP 80010, 67084 Strasbourg Cedex, [2] Collège Léon Dupriez, 1348, 1348 Ottignies-Louvain-la-Neuve, Belgium

[1] {a.kalinina, vassili, todiras}@unistra.fr, [2] thomas.francois@uclouvain.be

Abstract

This article presents a learner-oriented annotated lexical resource of French multiword expressions (MWEs) designed to support text adaptation in foreign language reading. MWEs, including idioms and collocations, pose major comprehension challenges for learners because their meaning is often non-compositional or depends on conventional lexical constraints. To address this issue, the study extends an existing verbal MWE database by integrating nominal and verbal MWEs annotated according to a linguistically grounded typology distinguishing idioms, opaque collocations, and transparent collocations. The resource was developed through a multi-step methodology combining automatic extraction from pedagogical corpora, manual annotation using decision-tree-based guidelines, and CEFR level assignment based on corpus distribution. The resulting dataset includes approximately 2,700 expressions enriched with detailed linguistic and learner-relevant metadata. Annotation campaigns involving native and non-native annotators showed moderate agreement, reflecting the gradient nature of phraseological opacity. By linking phraseological complexity with learner proficiency, this resource provides a reproducible framework for modeling MWE difficulty. It offers valuable support for text adaptation, readability assessment, and the development of NLP-based educational tools, contributing to improved accessibility of French texts for language learners.

Keywords: multiword expressions; foreign language reading; readability; learner-oriented lexical resource; decision-tree-based annotation

1. Introduction

Multiword expressions (MWEs)[1] constitute a persistent challenge for foreign language learners (Wray, 2002; Howarth, 1998). In French as a foreign language (FFL), learners frequently encounter expressions such as *tomber dans les pommes* ('to faint') or *course contre la montre* ('race against time') whose interpretation depends on phraseological conventions rather than purely compositional semantics. Because their meaning cannot always be inferred directly from their components and may rely on language-specific combining constraints (Nunberg et al., 1994; Mel'čuk, 1998), such expressions can hinder reading comprehension. More generally, figurative and phraseological language introduces additional interpretive difficulty, as learners must simultaneously decode unfamiliar vocabulary and recognize conventionalized multiword patterns in order to construct meaning (Boers, 2000; Siyanova-Chanturia & Martinez, 2015). Research in second language acquisition has shown that idiomatic and semi-idiomatic expressions are particularly resistant to acquisition (even for reception) and frequently lead to misinterpretation or avoidance strategies (Irujo, 1986, 287, Burger, 2007). Consequently, identifying and evaluating phraseological complexity is essential for improving the accessibility of pedagogical texts and for supporting automatic text adaptation in FFL contexts.

Meanwhile, advances in natural language processing (NLP) have created new opportunities for developing tools to support foreign language reading, such as readability assessment systems like TextEvaluator (Sheehan et al., 2014), FLELex-based readability tools (François et al., 2014) and Newsela (Nushi, 2020), and foreign

1 We use the term *multiword expression* (MWE) rather than the term *phraseological expression* frequently used in linguistics, because it is the standard terminology in computational linguistics and by the projects which produced annotated resources on which this work builds, such as PARSEME, UniDive, and PolyLexFLE. The term MWE provides a broad and operational category encompassing idioms and collocations, while remaining compatible with computational annotation frameworks.

Proceedings of the Joint Workshop on Readability and Text Simplification (READIxTSAR) @ LREC 2026, pages 181–192
11 May 2026.

language learning, e.g. through intelligent language-learning platforms such as SimpleApprenant (Todirascu et al., 2019) or Revita (Katinskaia et al., 2018), which provide lexical and phraseological support to learners. These systems can help identify expressions that may hinder comprehension. However, their effectiveness depends on the availability of annotated resources that capture fine-grained distinctions between types of MWEs, their CEFR level and explicitly encode degrees of semantic opacity. While several linguistically oriented MWE annotated corpora are available (cf. Savary et al., 2018; Savary et al., 2024), resources designed to support language learning applications remain relatively scarce, especially for French and when opacity is concerned.

This paper presents a learner-oriented annotated lexical resource of French nominal and verbal MWEs. Nominal MWEs include compound nouns and nominal collocations (e.g., *course contre la montre* 'race against time', *forte pluie* 'heavy rain'), while verbal MWEs include idiomatic and collocational verb-based constructions (e.g., *poser un lapin* 'to stand someone up', *prendre une decision* 'to take a decision').These two categories are treated within a unified typological framework based on compositionality and semantic opacity, ensuring consistency across syntactic types.

The resource combines a linguistically grounded typology for rich characterization of MWE, explicit annotation guidelines based on decision trees, and learner-relevant metadata, including CEFR levels. Besides the resource itself, we also aim to provide a reproducible framework for characterizing MWE complexity in the context of text adaptation and technology-enhanced reading in foreign language education, with a particular focus on the methodology for acquiring MWEs.

The remainder of this paper is structured as follows. Section 2 reviews the role of phraseological complexity in foreign language comprehension and presents the typology adopted in this work. Section 3 describes the methodology used to extend and annotate the lexical resource. Section 4 presents the corpus sources and data selection procedures, while section 5 details the annotation protocol, including the decision-tree framework and the annotation campaigns. Section 6 then describes the automatic projection of the annotated MWEs onto a learner corpus and Section 7 presents the subsequent contextual validation by human annotators. Finally, Section 8 discusses the limitations and future work and Section 9 concludes the paper.

2. MWE Complexity and Foreign Language Comprehension

This section develops the theoretical foundations underlying our approach to MWE complexity. We first examine phraseological complexity as a dimension of readability in foreign language learning. We then introduce the continuum-based typology of MWEs adopted in this study, which serves as the conceptual background for our annotation framework. Finally, we discuss the role of MWEs in pedagogical materials and their implications for text adaptation. Together, these elements clarify the linguistic and pedagogical motivations for the learner-oriented resource proposed in this article.

2.1 Phraseological complexity as a dimension of readability

Traditional approaches to readability focus on word length, lexical sophistication, and sentence length (Chall, 1996; Alderson, 2000). More recent work has broadened these approaches by introducing finer-grained lexical features, such as lexical diversity, frequency and familiarity; parse-based syntactic metrics; or various discourse properties which provide a more accurate account of text difficulty for language learners (Crossley et al., 2011; François & Fairon, 2012; Gooding et al., 2021). Phraseological complexity constitutes an additional dimension that is particularly salient in foreign language reading but have been hardly investigated within readability. Ozasa et al. (2007) presented an EFL readability formula for Japanese learners that includes, among other variables, an index of textbook-based idiom difficulty. However, this variable was not significant in its multiple linear regression model (Ozasa et al., 2007, 4). Later, François and Watrin (2011) assessed the contribution of various predictors based on MWE to readability formula to be close to negligible. However, they only considered nominal MWE, used an imperfect detection strategy, and were not able to distinguish opaque MWE. In line with this perspective, Kochmar et al. (2020) consider idioms as indices of text complexity. As illustrated in their results, idioms rank second in mean complexity, immediately after compounds, and clearly above several other MWE types such as verb-preposition constructions, coordinated phrases and semi-fixed expressions.

MWE should however have an impact on reading. For example, a sentence containing only frequent vocabulary may still be difficult if it includes an unfamiliar idiom composed of frequent tokens. Conversely, a text composed of transparent collocations may be easier to process despite higher lexical density. Encoding precisely phraseological information could therefore refine existing readability measures and provides a more nuanced picture of learner reading difficulties.

2.2 From the Continuum of Compositionality to MWE Typology

Research in phraseology generally assumes that MWEs form a continuum ranging from fully compositional combinations to fully idiomatic expressions (Sag et al., 2002:4). At one end are free combinations, whose meaning is predictable from that of their components. At the other end are idioms such as *tomber dans les pommes* (‘to faint’, literally ‘to fall into the apples’), whose figurative meaning cannot be inferred from the meanings of the individual words. Between these two extremes lie collocations (Tutin & Grossman, 2002; Grossmann & Tutin, 2003), which remain semantically transparent to varying degrees but are characterized by conventional lexical restrictions governing the choice of their components.

To operationalize this continuum in our work, we classify expressions according to a tripartite typology designed to capture compositionality (Mel’čuk, 1998) and graded semantic opacity (Gross, 1996):

Idiomatic expressions. These expressions are strongly non-compositional. Their global meaning cannot be inferred from the meanings of their components. Examples include *tomber dans les pommes* (‘to faint’) and *poser un lapin* (‘to stand someone up’). Moreover, they present specific invariant morphosyntactic properties, and lexical fixedness. Such expressions typically require explicit learning and are major sources of misunderstanding for language learners.

Opaque collocations. These expressions involve metaphorical or metonymic mechanisms that partially obscure the compositional interpretation. For instance, *course contre la montre* (‘race against time’) invokes a metaphorical mapping between time pressure and competition. The collocations are more flexible to morphosyntactic and syntactic modifications. Learners may grasp individual components while miss figurative relations.

Transparent collocations. These expressions remain semantically accessible but are constrained by conventional lexical choices. Examples include *signer un contrat* (‘to sign a contract’) and *forte pluie* (‘heavy rain’). Although their meaning is readily inferable, learners must acquire the target-like collocational patterns in order to use them appropriately in production.

For foreign language learners, this classification reflects increasing levels of interpretation difficulty, ranging from transparent collocations, whose meaning is largely accessible, to opaque collocations, and ultimately to idiomatic expressions, whose figurative meaning cannot be inferred from their components (Ellis, 2008; Conklin & Schmitt, 2012; Kochmar et al., 2020). Idiomatic expressions, such as *prendre la poudre d’escampette* (‘to run away quickly’), may require explicit instruction or repeated exposure because their figurative meaning is not directly inferable from their components. Opaque collocations, such as *célibataire endurci* (‘confirmed bachelor’), involve metaphorical or metonymic mappings that may not be shared across languages and can therefore complicate interpretation. Transparent collocations like *signer un contrat* (‘to sign a contract’) are generally semantically accessible, but they still require learners to acquire language-specific collocational properties (specific lexical associations, specific morpho-syntactic properties).

These distinctions are not merely theoretical: they have direct consequences for reading comprehension. When learners encounter an unfamiliar idiom in a text, they may interpret it literally or fail to integrate its intended meaning into the discourse. Opaque collocations can also hinder processing when learners are unable to readily access the underlying metaphorical or metonymic mappings. By contrast, transparent collocations are generally easier to infer in context, even though they may still pose challenges in other areas of language use, such as production or the selection of target-like collocational patterns (Gyllstad & Wolter, 2016). Thus, semantic transparency may serve as a key criterion for selecting and prioritizing MWEs according to their relevance and difficulty for language learners (Barghamadi et al., 2023).

2.3 Phraseology and pedagogical materials

Pedagogical texts for French L2 language learning inevitably contain a wide range of MWEs, including idioms, opaque and transparent collocations, reflecting the pervasive role of formulaic language in natural discourse (Erman & Warren, 2000). Teachers frequently adapt pedagogical texts by explaining, glossing, or reformulating MWEs that may hinder comprehension, a process that requires pedagogical judgement and familiarity with learner difficulties (Nation, 2009, 58-59; Boers, 2013, 213). While such adaptations are common practice in foreign language instruction, they are often time-consuming and depend heavily on individual expertise.

A systematic resource that identifies and encodes MWEs and their degree of opacity could make the adaptation of pedagogical texts more consistent. By combining CEFR level information with descriptions of interpretation difficulties, such a resource may indeed support teachers' pedagogical judgement in adapting texts. In practice, this information can guide decisions about text modification: idiomatic expressions may be retained with glosses or contextual support, whereas highly opaque expressions in beginner materials may require reformulation or explicit explanation.

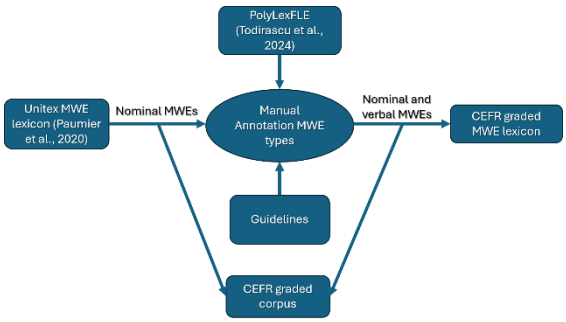


Figure1. Pipeline for the integration and CEFR level annotation of nominal and verbal MWEs in PolyLexFLE

3. The Method

In this study, we aim to extend the PolyLexFLE (Todirascu et al., 2024) lexical database by enriching it with nominal MWEs and providing consistent learner-oriented phraseological annotation across both nominal and verbal expressions. Our methodology combines corpus-based extraction, manual linguistic annotation, and proficiency-level assignment within a unified operational framework. Concretely, this extension process followed four main steps.

Firstly, candidate nominal MWEs were automatically identified by projecting external lexical resources such as the UniTex MWE lexicon (Paumier et al., 2020) onto CEFR-annotated pedagogical corpora. This procedure ensured that only expressions attested in learner-relevant input were considered and enabled their distribution across proficiency levels to be observed.

Secondly, candidate expressions were annotated according to the typology presented in Section 2, which distinguishes MWEs based on their degree of semantic compositionality and opacity (Section 5). Since semantic opacity has been shown to influence learner difficulty (Kochmar et al., 2020), the assigned category serves as an indicator of the expected processing complexity for learners. The initial annotation was deliberately carried out without full contextual embedding (while providing a usage example), in order to facilitate the identification and categorization of MWEs. This procedure allowed annotators to focus on the intrinsic properties of expressions (e.g., compositionality, lexical constraints) without missing candidates due to insufficient or ambiguous contextual cues and ensured stable lexical classification.

Thirdly, the newly annotated nominal MWEs were integrated with the existing verbal entries of PolyLexFLE (reannotated according to our typology), resulting in a unified dataset combining phraseological category and detailed annotation metadata, including traces of annotator reasoning and applied diagnostic tests.

Finally, the annotated inventory was integrated into a broader annotation pipeline. Firstly, the manually annotated MWEs were projected onto pedagogical corpora in order to identify their occurrences and assign their CEFR levels (Section 6). Secondly, the annotators reviewed automatic projections on the corpus, corrected errors, identified and annotated additional MWEs not present in the initial inventory (Section 7). These newly identified expressions were then incorporated into the PolyLexFLE database, allowing the resource to be progressively enriched. This two-step process complements the out-of-context annotation by introducing corpus based validation and improving both the accuracy and the coverage of the resource.

This methodology ensures that the resulting resource is simultaneously corpus-grounded,

linguistically motivated, and learner-oriented, providing a reproducible foundation for the study of phraseological complexity in foreign language learning.

4 Corpus sources and data selection

The extended dataset comprises approximately 2,700 (800 verbal and 1900 nominal) manually annotated MWEs extracted from pedagogical corpora representing instructional materials for learners of French as a foreign language.

Nominal MWEs were identified by projecting the Unitex compound lexicon (Paumier et al., 2020), filtered for nominal compounds, onto a corpus of FFL teaching materials assembled within the ANR STAR-FLE project[2]. This corpus includes 35 textbooks ranging from CEFR level A1 to C2 from which we extracted more than 400 learner-oriented texts (more than 500 000 tokens) covering a range of genres, such as dialogues, narratives, and informational texts. This procedure enabled the systematic identification of candidate expressions for manual annotation and CEFR-level assignment.

The verbal MWEs originate from the freely available PolyLexFLE database and are associated with CEFR proficiency levels obtained from two complementary sources. Firstly, the CEFR levels of several expressions were extracted from reference level vocabularies (Beacco, 2007; Beacco, 2008; Beacco & Porquier, 2008), providing expert-based level assignments. Secondly, additional expressions were automatically assigned CEFR levels using the lowest-level occurrence method in the corpus, relying on their distribution across CEFR-annotated pedagogical corpora compiled within the SimpleApprenant project (Todirascu et al., 2019, 2024).

5 The Annotation Campaign

5.1. Decision-tree-based annotation

A central contribution of this work is an operational annotation protocol based on explicit decision trees composed of linguistically motivated tests, partly inspired by the PARSEME framework (Savary et al., 2018) and UniDive linguistic diagnostics (Savary et al., 2024). The aim of the annotation is to classify the MWE according to the typology presented in section 2. We developed separate annotation guidelines for nominal and verbal MWEs, translating linguistic criteria into structured sequences of empirically applicable tests.

Formal diagnostics target morphosyntactic stability, including resistance to lexical substitution and internal variation. Annotators had to evaluate whether replacing a component preserves acceptability or meaning – for example, whether an idiom such as *prendre la poudre d'escampette* ('to run away quickly') tolerates lexical substitution (**prendre la **poussière** d'escampette*) or whether its internal structure allows morphosyntactic modification (*prendre la poudre d'escampette* → **prendre la poudre de l'escampette*). Additional tests examine whether syntactic transformations, such as passivation for verbal constructions, are possible (for example, *prendre la poudre d'escampette* → **la poudre d'escampette **a été prise***), or whether modifiers can be inserted without disrupting the idiomatic meaning (*étoile filante* 'shooting star' → **étoile **très** filante*). These diagnostics help determine the degree of structural rigidity and lexical fixedness that characterize MWEs.

Semantic diagnostics address degrees of compositionality and opacity and the role of figurative mechanisms. Annotators had to assess whether the global meaning can be inferred from the meanings of the components and whether metaphorical or metonymic mappings are central to interpretation. For instance, expressions such as *célibataire endurci* ('confirmed bachelor') require recognition of figurative associations that are not recoverable through literal composition: *endurci* (litt. 'hardened') is applied for raw materials, not for describing human beings.

5.2. Annotation protocol and training

Annotators received detailed guidelines and training sessions including several examples. For each expression, they applied the decision-tree tests and documented their reasoning directly in the resource, ensuring traceability of the annotation process.

Borderline cases were flagged for discussion, and the resulting changes were incorporated into the annotation guidelines. This iterative process allowed us to preserve explicit traces of the

2 https://anr.fr/Projet-ANR-23-CE38-0007

annotators' reasoning while refining and stabilizing the category boundaries.

Each expression was independently annotated by multiple participants. Majority decisions established reference labels, and unresolved cases were adjudicated by expert linguists. Recording the sequence of applied tests ensures transparency and supports subsequent analysis of disagreements.

5.3. Example annotation cases

Consider the expression *mettre la main à la pâte* ('to pitch in'). Annotators first apply the decision tree for verbal idiomatic expressions (see Figure 2). The expression passes the structural tests (MORPH and MORPHSYNT) and fails the compositionality test (COMP), since its global meaning cannot be derived from the literal meanings of its components. It is therefore clasified as an idiomatic expression.

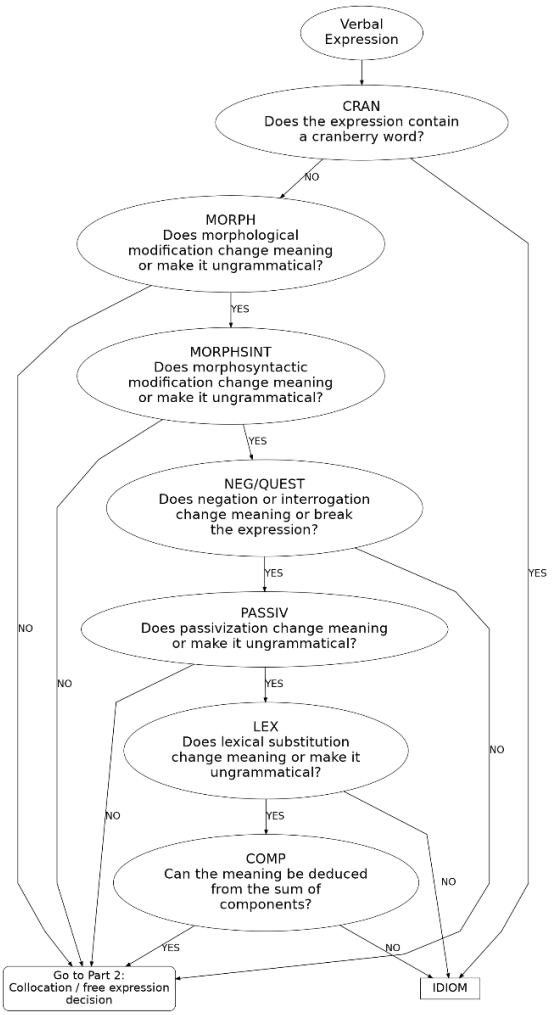


Figure 2. Part 1: Verbal idiom decision tree

By contrast, *prendre une décision* ('take a decision') is redirected from the decision tree for verbal idiomatic expressions to the decision tree for verbal collocations (see Figure 3).

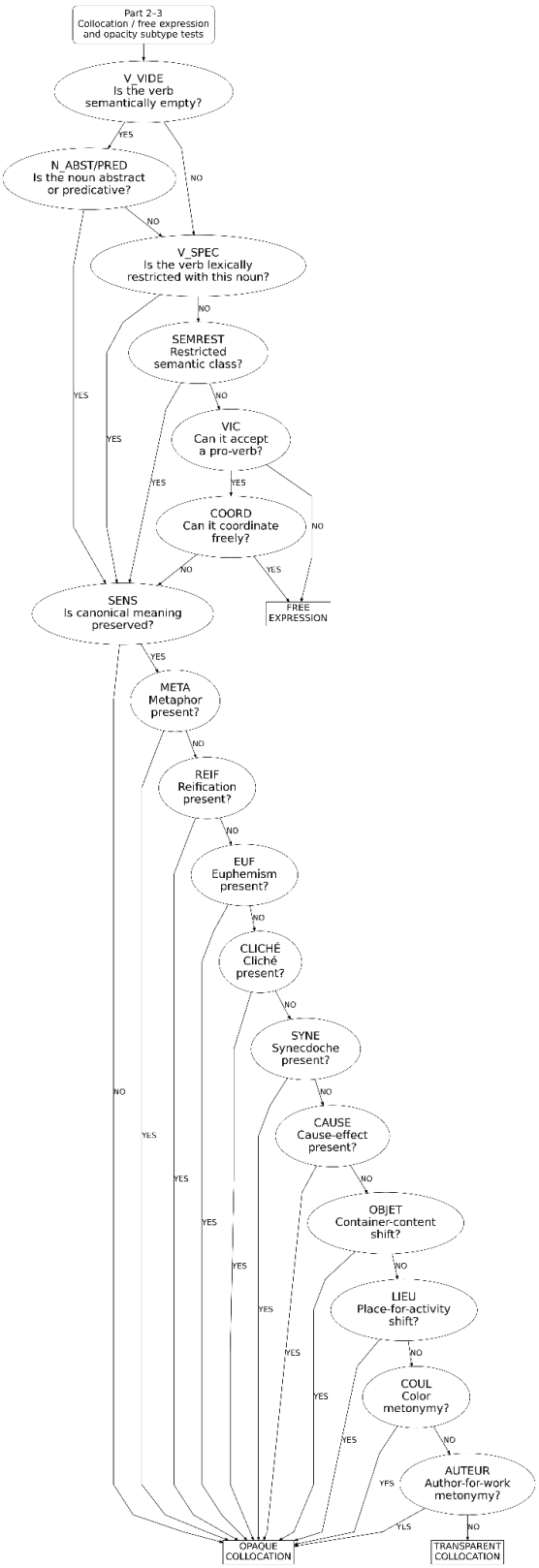


Figure 1. Part 2: Verbal collocation decision tree

The verb *prendre/take* is a semantically light verb (V_VIDE) combined with an abstract noun *décision* (N_ABSTR/PRED), confirming the collocational status of the MWE.

Opacity diagnostics further show that the expression involves reification (REIF), while an abstract process (*décision*) is conceptualized as a manipulable object. It is thus categorized as an opaque collocation.

Such step-by-step application of the annotation tests illustrates how the decision-tree framework implies phraseological distinctions in a transparent and reproducible way.

5.4 Inter-annotator agreement

Two MWE annotation campaigns were conducted to evaluate the robustness and reproducibility of the proposed framework. The first campaign involved advanced non-native speakers of French with linguistic training (master and Ph.D. students, senior researchers), while the second involved native speakers also with linguistic training (master students and senior researchers). This dual design made it possible to examine how annotator profiles influence MWE classification, particularly in borderline cases.

Each expression was first independently annotated out of context by three to ten annotators, who assigned it to one of the three classes described in Section 2. The number of annotators per expression varies due to the organization of annotation campaigns across multiple sessions. Expressions were distributed incrementally as part of training and evaluation phases, resulting in heterogeneous coverage. However, all expressions were annotated by at least three annotators, ensuring a minimal level of redundancy, while a subset of expressions received additional annotations to support finer-grained agreement analysis.

Inter-annotator agreement, measured for all annotators independent of their profile using Cohen's kappa, ranged from 0.32 to 0.39 (mean = 0.355) across annotator pairs. When compared to the Gold annotation – defined as the majority label for each expression – the kappa values of each annotator ranged from 0.43 to 0.62 (mean = 0.502). These scores fall within the moderate range and are consistent with previous findings that the annotation of gradient semantic phenomena poses substantial challenges (Artstein & Poesio, 2008). Nevertheless, the decision-tree framework provides a structured basis for annotation by operationalizing semantic and formal diagnostics into explicit classification steps and requiring annotators to document their reasoning. This approach promotes transparency, reproducibility, and systematic evaluation, which are recognized as key factors for ensuring reliable semantic annotation (Pustejovsky & Stubbs, 2012; Savary et al., 2018).

Agreement varies across categories. Idiomatic expressions show the highest convergence: annotators tend to agree when an expression clearly violates compositional interpretation. For example, expressions such as *tomber dans les pommes* or *poser un lapin* are consistently recognized as idiomatic because their figurative meaning sharply contrasts with literal interpretation. By contrast, the boundary between opaque and transparent collocations generates more disagreement. Expressions involving weak metaphorical extensions, such as *prix élevé* ('high price'), or partially conventionalized meanings, such as *vive émotion* ('intense emotion'), frequently trigger divergent judgments, confirming that semantic opacity constitutes a continuum rather than a binary distinction (Nunberg et al., 1994; Gibbs, 1994).

5.5. Sources of disagreement

A qualitative analysis of disagreements reveals several recurring patterns.

Firstly, annotators differ in their sensitivity to metaphorical interpretation. Some classify expressions with faint figurative traces as opaque, while others emphasize compositional accessibility and assign them to transparent collocations. Expressions encoding abstract processes via spatial imagery, such as *entrer en vigueur* ('enter into force'), often lie at the boundary between opaque and transparent collocations, depending on the extent to which annotators perceive the underlying metaphorical mapping.

Secondly, annotator profile influences annotation behavior in ways that are not reducible to simple cross-linguistic interference. In our experiments, non-native annotators often adhered more closely to the explicit decision-tree tests than native speakers, who tended to rely more heavily on intuitive judgments of naturalness. Because native speakers process familiar expressions holistically, they may be less sensitive to the formal and semantic diagnostics required by the protocol. By contrast, non-native annotators, accustomed to analytic processing of

phraseological material, tend to apply the tests more systematically. This asymmetry highlights the importance of considering annotator background when designing learner-oriented annotation frameworks.

A quantitative comparison further highlights differences between annotator groups. Native annotators show a higher average agreement ($\kappa \approx 0.35$) with relatively limited dispersion, suggesting more homogeneous judgments, likely driven by shared linguistic intuitions. In contrast, non-native annotators exhibit a lower average agreement ($\kappa \approx 0.27$) but greater variability across annotators. This dispersion indicates more heterogeneous annotation strategies: while some non-native annotators closely follow the decision-tree guidelines, others show greater uncertainty when evaluating semantic opacity. These results suggest that annotator background influences not only agreement levels but also the balance between intuitive and analytical processing in MWE classification.

Thirdly, annotation decisions are shaped by how annotators interpret the contextual information provided with each expression. During the initial annotation stage, candidate expressions are presented in a structured spreadsheet together with an explicit usage example specifying the intended reading. Although this controlled contextualization reduces ambiguity, some expressions remain compatible with alternative interpretations along the compositionality continuum. Differences in how annotators balance the provided contextual cue against their lexical intuitions therefore introduce an additional source of variability.

5.6 Role of the decision-tree framework

Despite these challenges, the decision-tree methodology proves effective in structuring annotator reasoning for such a complex annotation task. By requiring explicit evaluation of formal and semantic diagnostics, the framework limits purely intuitive judgments – which do not always lead to the most consistent annotations – and promotes systematic comparison across expressions. Annotators report that the ordered sequence of tests is particularly helpful in clarifying borderline cases.

The traceability of annotation decisions also supports iterative refinement of the guidelines. Clusters of disagreement reveal areas where additional examples or clarifications are needed. In this sense, the annotation campaigns function not only as evaluation tools but also as feedback mechanisms for improving the annotation protocol.

6. Automated projection of manual annotation onto an FFL corpus

After completing the out-of-context annotation of approximately 2,700 expressions, the manually validated inventory was automatically projected onto a corpus of FFL textbooks and assessment materials. This projection was performed using the Stanza annotation pipeline in order to identify corpus occurrences of the annotated MWEs and to construct a phraseologically annotated learner corpus.

In addition to marking expression occurrences in context, this step enabled the attribution of CEFR levels to each MWE following the methodology proposed by François et al. (2014) and Todirascu et al. (2024). Each expression was assigned the lowest CEFR level of the pedagogical material in which it appeared, thereby linking phraseological items to empirically attested instructional contexts. An analysis of CEFR-level distribution shows that transparent collocations are more frequent at lower proficiency levels (A1–B1), although specialized or terminological collocations are attested at higher levels up to C2. In contrast, opaque collocations and idiomatic expressions become more frequent at higher levels (B2–C2). This distribution aligns with pedagogical expectations regarding the gradual introduction of phraseological complexity.

The result of the automated projection process is a semi-automatically annotated corpus that combines lexicon-based projection with contextual occurrence data. This corpus serves both as a resource for studying the distribution of MWEs in pedagogical materials and as an intermediate layer for subsequent human validation.

7. Validation of automatic annotation in context by human annotators

The automatically annotated corpus was then imported into the INCEpTION annotation platform for human validation. Annotators were asked to review the projected annotations in context and to perform three types of operations: validating correct automatic annotations, correcting misidentified expressions, and annotating missing MWEs that were not captured during automatic

projection, for instance because they were absent from the initial lexical inventory.

This second annotation phase introduces contextualized validation into the workflow, complementing the initial out-of-context categorization. Working with full textual context allows annotators to assess how MWEs function in authentic pedagogical materials and to refine the resource accordingly. The interaction between automatic projection and human validation thus creates an iterative annotation pipeline in which lexical annotation and corpus annotation mutually inform each other, strengthening both coverage and reliability of the final resource.

Preliminary results from this validation stage show that fewer than half of the MWEs manually identified in the corpus were captured by the automatic pre-annotation, which relied on the Unitex lexicon (Paumier et al., 2020) and the PolyLexFLE database (Todirascu et al., 2024). These results suggest that the recall of automatic pre-annotation remains below 50%, confirming the limitations of lexicon-based detection. However, the majority of automatically identified MWEs (24 out of 26 in the test text) were confirmed as valid expressions upon validation, indicating relatively high precision (91.66% for the test text). This result underscores the complementary roles of automatic extraction and manual validation. This discrepancy is particularly frequent for verbal MWEs, which are more than three times less numerous in the current lexical inventory than nominal ones. These findings make clear the limitations of lexicon-based automatic projection when applied to authentic pedagogical corpora and underscore the essential role of manual, context-based validation. In particular, human annotation in context is necessary to identify expressions that are absent from existing resources, exhibit contextual variation, or fall outside predefined lexical inventories. This validation phase therefore plays a crucial role in improving both the coverage and the representativeness of the final resource, ensuring that it more accurately reflects the diversity of MWEs encountered by language learners in real instructional materials.

8. Limitations and future work

Although the dataset covers a substantial number of expressions, its current scope remains limited to selected pedagogical corpora and specific categories of MWEs. Extending its coverage to additional genres and phraseological types would improve its representativeness and enable a more comprehensive account of phraseological complexity. In addition, the moderate inter-annotator agreement reflects the inherently gradient nature of phraseological phenomena, suggesting that future work may explore alternative representational approaches that better capture this variability. In particular, semantic opacity could be represented as a continuous scale rather than discrete categories. Disagreement patterns themselves may also provide useful information for identifying borderline cases.

Another important direction for future research is to evaluate the pedagogical relevance of the proposed typology with actual learners of French as a foreign language. In particular, empirical studies could investigate whether the distinctions between idiomatic expressions, opaque collocations, and transparent collocations correspond to differences in learners' comprehension difficulty and processing, thereby validating the learner-oriented adequacy of the framework.

Beyond its descriptive contribution, the resource offers promising pedagogical and technological applications. It may support text adaptation and the development of learner-oriented materials by helping identify expressions that are likely to challenge comprehension. Its structured format also makes it suitable for integration into NLP-based educational tools, enabling automatic detection and pedagogical support for MWEs.

The dataset will be made publicly available upon publication in order to support reproducibility and further research. It will be distributed in a structured format including MWE entries, typological annotations, CEFR levels, and annotation metadata.

9. Conclusion

We present a learner-oriented annotated resource of French multiword expressions designed to support text adaptation in foreign language reading. The resource combines a linguistically grounded typology, explicit decision-tree annotation guidelines, and learner-relevant metadata. Annotation campaigns demonstrate moderate agreement and highlight the intrinsic challenges of modeling graded semantic opacity.

By systematically encoding phraseological compositionality and opacity for language learners, the resource contributes to bridging

linguistic theory, NLP, and foreign language pedagogy. It supports both manual and computational approaches to identifying expressions that hinder comprehension and provides a foundation for technology-enhanced reading tools.

More broadly, this work emphasizes the importance of integrating phraseological knowledge into models of automatic readability assessment and automatic text adaptation. Future research will extend the resource and explore its application in adaptive educational technologies. The resource is also compatible with recent large language model (LLM)-based approaches, which opens the way for systematic comparisons between human annotations and LLM predictions. MWE annotations could be used as control signals for simplification, as evaluation benchmarks for phraseological processing, or as supervision data for fine-tuning models to better handle phraseological complexity. In addition, future developments may include the introduction of graded annotations, reflecting degrees of opacity or annotator confidence, in order to better capture the continuum of phraseological phenomena.

Through these efforts, we aim to improve the accessibility of foreign language texts and support learners in navigating the rich phraseological landscape of French.

10. Acknowledgements

This research was conducted within the framework of the ANR STAR-FLE project, whose support is gratefully acknowledged. The authors also wish to express their sincere thanks to CENTAL for providing financial support for the annotation campaign.

Thomas François is supported by the Belgian FNRS through the action PDR 40013622.

11. Lay Summary

This study presents a resource designed to help people learning French to better understand expressions composed of several words. These expressions, such as idioms or common word combinations, might be difficult because their meaning is not always clear from the individual words or depends on knowing typical usage patterns.

To address this issue, we expanded an existing lexical database by adding both verb-based and noun-based expressions. Each expression is classified according to how easy it is to understand: some are easy to understand some are partly figurative, and others cannot be understood literally at all.

The resource was created through several steps. First, expressions were automatically identified in language learning materials. Then, trained annotators analyzed and classified them using explicit guidelines. Finally, each expression was related to a learner level (from beginner to advanced) based on where it occurs in teaching texts.

The final dataset contains around 2,700 expressions, along with useful information about their structure and difficulty for learners. The study also shows that even experts do not always fully agree on how to classify these expressions, which reflects the fact that their meaning can be more or less transparent.

By connecting the difficulty of these expressions to learner levels, this resource can help teachers, researchers, and digital tools better identify what may be hard to understand in a text. It can support text simplification, improve readability assessment, and contribute to the development of tools that make French texts more accessible for learners.

12. Bibliographical References

Alderson, J. C. (2000). *Assessing reading*. Cambridge: Cambridge University Press.

Artstein, R. & Poesio M. (2008). Survey Article: Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

Barghamadi, M., Rogers, J., Arciuli, J., Müller, A. (2023). The use of semantic transparency and L1-L2 congruency as multi-word units selection criteria. *Studies in English Language and Education*, 10(2):723–740.

Beacco, J.-C. & Porquier, R. (2008). *Niveau A2 pour le français : utilisateur-apprenant élémentaire*, Didier, Paris.

Beacco, J.-C. (2008). *Niveau A1/A2 pour le français: Textes et références.* Didier. Paris.

Beacco, J.-C., & Porquier, R. (2007). Niveau A1 pour le français: utilisateur-apprenant élémentaire. Didier. Paris.

Boers, F. (2000). Metaphor awareness and vocabulary retention. Applied Linguistics – APPL LINGUIST, 21:553–571.

Boers F. (2013). Cognitive Linguistic approaches to teaching vocabulary: Assessment and

integration. *Language Teaching*. 46(2):208–224.

Burger, H. (2007) (Ed.): Phraseologie. Ein internationales Handbuch der zeitgenössischen Forschung. Walter de Gruyter. Berlin.

Chall, J. S. (1996). Varying Approaches to Readability Measurement. *Revue québécoise de linguistique*, *25*(1):23–40.

Conklin K, Schmitt N. (2012). The Processing of Formulaic Language. *Annual Review of Applied Linguistics*. 2012; 32:45–61.

Crossley S. A., David A. & McNamara, D. (2011). Text readability and intuitive simplification: A comparison of readability formulas. *Reading in a Foreign Language*, 23:84–101.

Ellis, N. C. (2008). Phraseology: The periphery and the heart of language. *Applied Linguistics*, 29(1):1–13.

Francois, T., & Watrin, P. (2011). On the contribution of MWE-based features to a readability formula for French as a foreign language. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 441–447, Hissar, Bulgaria. Association for Computational Linguistics.

François T. & Fairon C. (2012). An "AI readability" Formula for French as a Foreign Language. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 466–477, Jeju Island, Korea. Association for Computational Linguistics.

François T., Gala N., Watrin P., Fairon C. (2014). FLELex: a graded lexical resource for French foreign learners. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3766–3773, Reykjavik, Iceland. European Language Resources Association (ELRA).

Gibbs, R. W. (1994). *The Poetics of Mind: Figurative Thought, Language, and Understanding*. Cambridge: Cambridge University Press.

Godwin-Jones, R. (2023). Emerging spaces for language learning: AI bots, ambient intelligence, and the metaverse. *Language Learning & Technology*, 27(2):6–27.

Gooding S., Berzak Y., Mak T., and Sharifi M. (2021). Predicting Text Readability from Scrolling Interactions. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 380–390, Online. Association for Computational Linguistics.

Gross, G. (1996). *Les expressions figées en français : Noms composés et autres locutions*. Ophrys.

Grossmann, F., Tutin, A. (Dir.). (2003). *Les collocations : analyse et traitement.* De Werelt.

Gyllstad, H., & Wolter, B. (2016). Collocational processing in light of the phraseological continuum model: Does semantic transparency matter? *Language Learning*, 66(2):296–323.

Heift, T., & Schulze, M. (2007). *Errors and Intelligence in CALL. Parsers and Pedagogues.* New York: Routledge.

Howarth, P. (1998). Phraseology and second language proficiency. *Applied Linguistics*, 19(1):24–44

Irujo, S. (1986). Don't put your leg in your mouth: Transfer in the acquisition of idioms in a second language. *TESOL Quarterly*, 20(2):287–304.

Katinskaia A., Nouri J., and Yangarber R. (2018). Revita: a Language-learning Platform at the Intersection of ITS and CALL. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 4084–4093, Miyazaki, Japan. European Language Resources Association (ELRA).

Kochmar, E., Gooding, S. and Shardlow. M. (2020). Detecting Multiword Expression Type Helps Lexical Complexity Assessment. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4426–4435, Marseille, France. European Language Resources Association.

Mel'čuk, I. (1998). Collocations and Lexical Functions. In A.P. Cowie (ed.), *Phraseology. Theory, Analysis, and Applications*, Oxford: Clarendon Press, 23–5.

Nation, I.S.P. (2008). *Teaching ESL/EFL Reading and Writing*. Routledge. 1st edition.

Nushi, M. (2020). Newsela: A Level-Adaptive App to Improve Reading Ability. *Reading in a Foreign Language*. 32:239–247.

Nunberg, G., Sag, I. A., & Wasow, T. (1994). Idioms. *Language*, 70(3):491–538.

Ozasa, T., Weir, G., & Fukui, M. (2007). Measuring readability for Japanese learners of English. In *Proceedings of the 12th Conference of Pan-Pacific Association of Applied Linguistics;* pages 122-125, Pattaya, Thailand, December. Pan-Pacific Association of Applied Linguistics

Pustejovsky, J., & Stubbs, A. (2013). *Natural language annotation for machine learning: A guide to corpus-building for applications.* O'Reilly Media.

Sag, I.A., Baldwin, T., Bond, F., Copestake, A., Flickinger, D. (2002). Multiword Expressions: A Pain in the Neck for NLP. In: Gelbukh, A. (eds) Computational Linguistics and Intelligent Text Processing. CICLing 2002. *Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg, vol 2276:1–15.

Savary, A., Candito, M., Barbu Mititelu, V., Bejček, E., Cap, F., et al. (2018). PARSEME multilingual corpus of verbal multiword expressions. In *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*.

Savary, A., Zeman, D., Barbu Mititelu, V., Barreiro, A., Caftanatov, O., de Marneffe, M.-C., Dobrovoljc, K., Eryiğit, G., Giouli, V., Guillaume, B., Markantonatou, S., Melnik, N., Nivre, J., Ojha, A. K., Ramisch, C., Walsh, A., Wójtowicz, B., & Wróblewska, A. (2024). UniDive: A COST Action on universality, diversity and idiosyncrasy in language technology. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 372–382. ELRA & ICCL.

Sheehan, K. & Kostin, I. & Napolitano, D. & Flor, M. (2014). The TextEvaluator Tool: Helping Teachers and Test Developers Select Texts for Use in Instruction and Assessment. *The Elementary School Journal*, 115:184–209.

Siyanova, A. & Martinez, R. (2015). The idiom principle revisited. Applied Linguistics, 36(5):549–569.

Todirascu, A., & Cargill, M. (2019). SimpleApprenant: A platform to improve French L2 learners' knowledge of multiword expressions. In *Proceedings of EUROCALL 2019. Université catholique de Louvain & University of Leuven*, Louvain-la-Neuve, Belgique.

Todirascu, A., François, T., & Cargill, M. (2024). PolyLexFLE: A MWE database for French L2 language learners. *International Journal of Applied Linguistics*, 175(1):77–102.

Tutin, A. & Grossmann, F. (2002). Collocations régulières et irrégulières : esquisse de typologie du phénomène collocatif. *Revue française de linguistique appliquée*. VII(1):7–25.

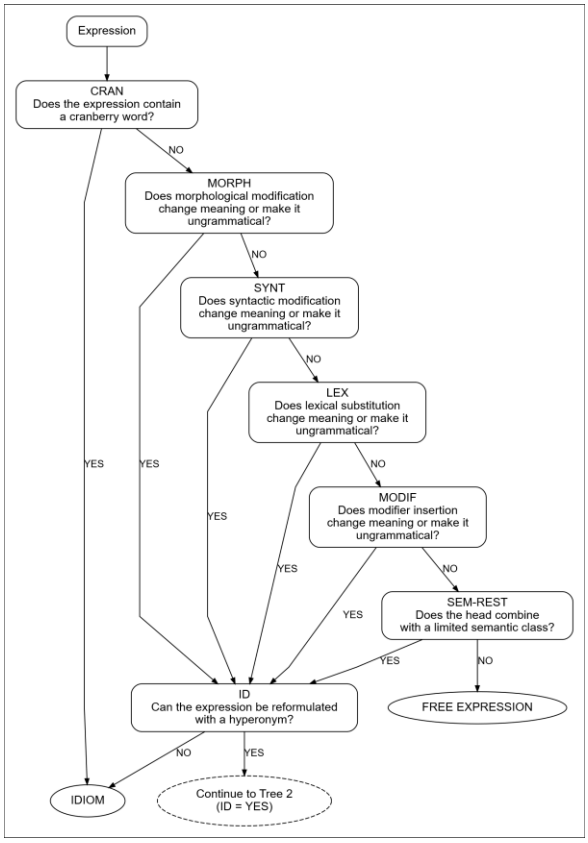
Wray, A. (2002). *Formulaic Language and the Lexicon*. Cambridge University Press.

13. Language Resource References

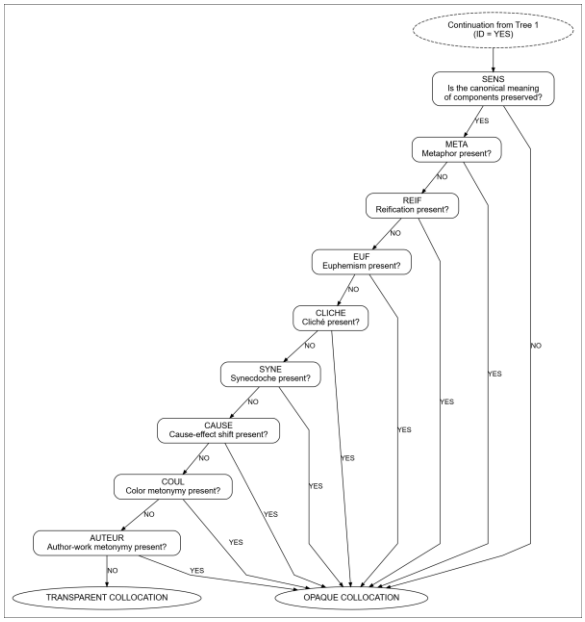
Paumier S. (2020). Unitex 3.2. Manuel d'utilisation, Université Paris-Est-Marne-la-Vallée, https://unitexgramlab.org/releases/3.2/man/Unitex-GramLab-3.2-usermanual-fr.pdf

.

Appendix A: Decision tree for nominal MWEs (Part 1: Idiom detection)



Appendix B: Decision tree for nominal MWEs (Part 2: Opacity detection)



A Meta-evaluation of Automatic Metrics for Elaborative Simplification

Abdullah Alshatti, Steven Schockaert, Fernando Alva-Manchego
School of Computer Science and Informatics, Cardiff University, UK
{AlshattiAM, SchockaertS1, AlvaManchegoF}@cardiff.ac.uk

Abstract

Elaborative simplification aims to improve the readability of texts by adding content that helps the readers. However, evaluating these elaborations remains challenging due to their subjective nature and the lack of suitable annotated datasets. To support the evaluation of elaborative simplification models, we introduce a new dataset with human ratings of elaborations generated by Large Language Models (LLMs), focusing on two quality criteria: cohesion and informativeness. Using these human judgments as a reference, we conduct a meta-evaluation of existing automatic evaluation approaches, with a focus on LLM-as-a-judge strategies. Our experiments suggest that evaluations made by smaller LLMs correlate poorly with human judgments, while larger models with structured prompting exhibit higher agreement. Informativeness evaluation proved to be challenging due to its subjectivity, as evidenced by the low inter-annotator agreement compared to cohesion.

Keywords: Text Simplification, Elaboration, Text Evaluation

1. Introduction

Elaborative simplification improves readability by adding explanations or information aligned with the original text's meaning and context (Srikanth and Li, 2021). Table 1 illustrates such an elaboration. The quality of elaborations is crucial, as inaccuracies or poorly worded extra information may render the text nonsensical and harder to understand (Long and Ross, 1993; Shardlow, 2014).

However, a critical gap exists in evaluating elaborative simplification, as traditional reference-based metrics have proven inadequate due to their reliance on lexical overlap with fixed references (Srikanth and Li, 2021). Even embedding-based metrics such as BERTscore (Zhang et al., 2020) are often poorly correlated with human judgments of quality (Moramarco et al., 2022; Li et al., 2024; Kryscinski et al., 2020; Fabbri et al., 2021). Overall, while traditional automated metrics offer scalability and objectivity, often fall short in generative tasks.

A potential solution is to rely on LLM-as-a-judge approaches, where the quality of elaborations is assessed by prompting an LLM, eschewing the need for reference answers. However, there are currently no annotated datasets that can be used for evaluating the reliability of LLM judges in elaborative simplification, nor is there a standardized framework that can be used for designing suitable rubrics to assist them.

In this paper, we address this gap by introducing ElabEval, a manually annotated dataset of elaboration quality.[1] Specifically, we collect human quality ratings for LLM-generated elaborations for anchor sentences from the ElabQUD dataset (Wu et al., 2023). Our annotations cover two aspects of quality: (i) *cohesion*, measuring the extent to which an elaboration is sensible within the given context, and (ii) *informativeness*, measuring how useful the provided information is likely to be to the reader.

CONTEXT: Anderson became interested in people like Landa when she noticed something strange about a call center near her house.
ELABORATION: Workers at call centers help people over the phone.

Table 1: Example of an elaboration from ElabQUD (Wu et al., 2023), answering the implicit question under discussion: *What do call centers do?*

Using this dataset, we conduct a meta-evaluation of standard reference-based metrics and reference-free LLM-as-a-judge approaches. We also leverage the LLM-as-a-Judge approach to provide a scalable alternative to human assessment and compare its reliability against both traditional automatic metrics and human judgments, employing a variety of LLMs to evaluate elaborations, ensuring a more comprehensive assessment across model sizes and architectures. Our analysis confirms that standard metrics, which rely on comparing the generated elaborations with reference elaborations from the ElabQUD dataset, perform poorly. When it comes to (reference-free) LLM judges, we found smaller models to perform surprisingly poorly, barely outperforming random guessing. However, we also found that frontier models such as gpt5 can provide reliable assessments when sufficiently detailed instructions are provided.

Overall, our results show that the proposed

[1] Resources available on: https://github.com/Abdullah-alshatti/ElabEval

Proceedings of the Joint Workshop on Readability and Text Simplification (READIxTSAR) @ LREC 2026, pages 193–209
11 May 2026.

dataset provides a realistic reference for meta-evaluating automatic evaluation methods for elaborative simplification, while also revealing the current limitations of LLM-as-a-judge approaches.

2. Related Work

Datasets. Srikanth and Li (2021) introduced the term "elaborative simplification" to describe content addition in text simplification to improve readability. Through crowdsourcing, they collected a dataset of 1.3K naturally occurring elaborations in the Newsela corpus (Xu et al., 2015) focusing on the contextual aspect for these elaborations. Based on this annotated dataset, Wu et al. (2023) used human annotation and encoder-decoder models to generate an implicit Question Under Discussion (QUD) to help guide LLMs in producing contextually relevant elaborations. Laban et al. (2023) created the SWIPE dataset, which reconstructs the document-level editing process from English Wikipedia (EW) articles to paired Simple Wikipedia (SEW) articles by leveraging the entire revision history during the pairing process in order to better identify simplification edits. In addition to other types of simplifications, this dataset contained instances of elaboration as well.

Since the quality of Wikipedia-based datasets in simplification is questioned (Trokhymovych et al., 2024), we opted for ElabQUD, which is the Newsela corpus, a professionally curated resource of English-language texts. This choice ensured higher data quality while also introducing QUD-based elaboration structures that enriched the diversity of our dataset.

Methods. Existing approaches to generate elaborative simplifications mainly use Transformer (Vaswani et al., 2017) based models , relying on prompting and fine-tuning of pre-trained language models. For example, Srikanth and Li (2021) fine-tuned GPT-2 (Radford et al., 2019), using the simplest texts in Newsela and their annotated elaborations, then provided the model with the text preceding the elaboration in a simplified text as input, and the model would generate the elaboration as output. Wu et al. (2023) used GPT-3 (Brown et al., 2020) for zero-shot elaboration generation, experimenting with including an automatically generated and manually written QUD in the prompt, finding that the latter type produced the best elaborations. For German, Hewett et al. (2024) used Meta-Llama-3-8B-Instruct (Grattafiori et al., 2024) as both out-of-the-box model and fine-tuned on B1 and A2 German texts, following the CEFR language proficiency framework. They also used prompt variations to generate elaborations with generic, background, and contextual information.

These approaches are also used in definition generation, a related task. Yarbro and Olney (2021) used a dataset containing words definitions and a list of contexts associated them to fine-tune a GPT-2 based model to generate definitions for English words with only the word and a context as inputs. Asthana et al. (2024) used four LLMs: GPT-4 (OpenAI, 2024), PaLM-2 (Anil and et al, 2023), Falcon-40b (Almazrouei et al., 2023), and BLOOM-176b (Workshop et al., 2023). They provided the models with the term, definition, and difficult concept and used two types of prompts that reflect two simplification strategies, to rewrite the definition by adding an explanation for the difficult concept and to rewrite the definition and simplify the difficult concept word.

We adopted an approach similar to these works to generate elaboration instances for our dataset. We provide the LLM models with the contexts that need to be elaborated on, using two types of prompting approaches: QUD-based prompting and a descriptive prompt without specifying the elaboration type that needs to be generated.

Evaluation. Previous work on elaborative simplification (Wu et al., 2023; Laban et al., 2023) reported automatic metrics scores like BERTScore (Zhang et al., 2020), BLEU (Papineni et al., 2002) and SARI (Xu et al., 2016) for completeness but relied on humans to evaluate the elaborations through comparing various elaborations based on how coherent and elaboration-like they are. These works did not provide clear definitions of what that means, making the evaluations vague, unspecified, and reliant on the interpretation of the evaluators.

Existing text simplification metrics face significant limitations when evaluating elaborative simplification compared to standard simplification; current metrics either aggregate meaning preservation and simplification into single overall scores or target only meaning preservation, confirming that no single automatic metric captures all necessary evaluation criteria (Cripwell et al., 2024; Alfear et al., 2024; Guo et al., 2024; Alva-Manchego et al., 2021). Such approaches penalize novel content generation, resulting in poor correlation with human judgments (Li et al., 2024; Barayan et al., 2025).

Studies in Natural Language Generation tasks have found the LLM-as-a-judge approach can achieve high correlation with human judgments (Wang et al., 2025), outperforming traditional automatic metrics in this respect (Nguyen et al., 2024; Chen et al., 2023). We also found the LLM-as-a-judge approach to be suitable for evaluating elaborations, as they have an extensive output space and cannot be fully captured by a fixed number of references. However, for this to work, appropriate criteria for evaluating the quality of the elaborations need to be established first. We introduce

an explicit evaluation framework for elaborative simplification that evaluates the elaborations via a two-stage rubric: annotators first make a binary cohesion judgment (filtering out irrelevant, inconsistent, or merely repetitive elaborations) and then rate only cohesive outputs for informativeness on a three-level scale.

3. The ElabEval Dataset

This section describes the curation of ElabEval, our annotated English dataset for the *meta-evaluation* of elaboration quality, i.e. for assessing the reliability of elaboration evaluation metrics.

3.1. Curation of Elaborations

Source Texts. We selected 100 news articles from ElabQUD (Wu et al., 2023) to use as contexts for the elaborations in our dataset. In ElabQUD, for each context, both an implicit Question Under Discussion (QUD) and an elaboration are provided. As context for each elaboration, we considered up to 5 sentences prior to the elaboration, following the same setup as Srikanth and Li (2021). The source text underwent a preprocessing step that involved removing the special characters to provide the context as a single clean input to the LLMs.[2]

LLM-Generated Elaborations. To obtain a dataset covering a wide range of elaboration quality, we first observed that smaller LLMs produce highly variable outputs. Based on this, we selected two such models to generate elaborations for our dataset: DeepSeek-R1-Distill-Qwen-7B[3] (Guo et al., 2025) and FLAN-T5_Instruct-Mistral7B[4] (Jiang et al., 2023). These models were chosen because they are open-weight, supporting reproducibility, and they consistently produced a mix of high- and low-quality elaborations. Both models were accessed via the Hugging Face Transformers library. They were used with all parameters set to default values, and the maximum generated elaboration length was capped at 100 tokens to ensure comparability across the models. For each context, we generated four elaborations (one for each prompt type and model). For each source context, the LLMs were instructed to generate an elaboration using the following two prompts:

- **Descriptive Prompt:** We provide specific instructions for the elaboration generation task with a target length of one sentence. The prompt was: *"You are tasked with generating a brief elaboration of one sentence from a given context. The context will be in the form of a paragraph consisting of multiple sentences. You're required to generate an appropriate new sentence that adds to the reader's understanding of the context in a clear and coherent way. Ensure the new sentence is concise and directly relevant to the information presented. Context: {sentence} Answer: "*
- **QUD:** We mirrored the methodology of (Wu et al., 2023), creators of ElabQUD. Specifically, the model was provided with inputs structured in the following format: *"Context: < context >, Question: < question >, Answer:"*

3.2. Collecting Human Judgments

Concerns have been raised regarding the declining quality of outputs generated by widely used crowdsourcing platforms, such as Amazon Mechanical Turk (MTurk) (Chmielewski and Kucker, 2020). In addition, several specialized services, including FigureEight, have become unavailable for academic research following their acquisition by commercial entities (Gilardi et al., 2023). These concerns led us toward the recruitment of annotators with verified competencies. Although this strategy enhanced the reliability of the resulting data, it simultaneously imposed constraints on the attainable scale of the dataset.

Annotators. We recruited three native English-speaking evaluators: two postgraduate students with backgrounds in linguistics and a non-academic staff member. They were selected from a pool of 11 candidates based on a qualification task that assessed their understanding of the annotation guidelines (available in Appendix A).

Annotation Criteria. We focused on two primary criteria to assess elaborations: cohesion and informativeness. *Cohesion* is meant to be more objective, focusing on whether a given elaboration "makes sense", and is thus evaluated as a binary property. More precisely, for an elaboration to be cohesive, it should:

1. maintain relevance to the given context;
2. be free from logical inconsistencies;
3. not contain any misleading examples; and
4. not merely repeat parts of the original text.

Informativeness assesses the utility of the provided information, relative to the given context. Given

[2] The QUD framework views each sentence as the answer to an implicit or explicit question from prior context.

[3] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-7B

[4] https://huggingface.co/SanketAI/FLAN-T5_instruct-mistral7b

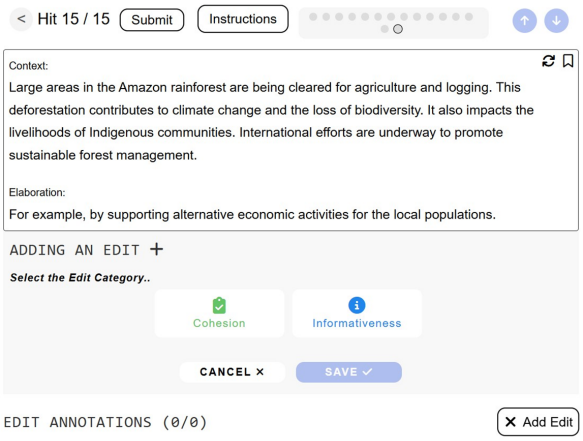


Figure 1: Screenshot of the annotation tool interface

the subjective nature of this criterion, it was evaluated using a three-point Likert scale. In particular, an elaboration is said to be *uninformative* if it offers no useful information for understanding the context, *somewhat informative* if it provides basic information that might benefit some readers, and *informative* if it would benefit most readers (e.g. by providing in-depth perspectives).

Annotation Tool. To collect the human ratings, we used the "thresh" annotation tool (Heineman et al., 2023), which is a customizable open-source platform for textual annotation.[5] Figure 1 shows a screenshot of our annotation interface created using this tool.

Annotation Process. Each annotator evaluated 500 elaborations: 100 gold-standard elaborations from ElabQUD and 400 generated by LLMs. For each instance, annotators first provided a cohesion judgement. Since incohesive elaborations are not informative, only those judged to be cohesive by an annotator were subsequently assessed for informativeness by the same annotator. As a result, informativeness ratings are only available for 226 out of the overall 500 instances: 165 have three evaluations for informativeness, 61 have two evaluations, and 5 instances have just one. In our analysis of informativeness, we only focus on those instances with three informativeness annotations. We convert the three ratings to a single value by summing them. By interpreting the individual ratings numerically, on a scale from 1 to 3, we thus obtain an overall informativeness score between 3 and 9. For cohesion, the final label was obtained via majority voting based on the scores given by the annotators.

[5] https://github.com/davidheineman/thresh

System		Level 1	Level 2	Level 3
Golden Elaborations		18	52	25
T5-Mistral	(DP)	8	2	0
	(QUD)	12	13	6
DeepSeek-R1-Qwen	(DP)	9	15	13
	(QUD)	18	28	7

Table 2: Distribution of informativeness labels (Level 1: Uninformative, Level 2: Somewhat Informative, Level 3: Informative) for the golden elaborations, the systems and prompting methods used: (DP) Descriptive Prompt and (QUD) Questions Under Discussion.

System		Cohesive	Incohesive
Golden Elaborations		95	5
T5-Mistral	(DP)	10	90
	(QUD)	31	69
DeepSeek-R1-Qwen	(DP)	37	63
	(QUD)	53	47

Table 3: Distribution of cohesion labels for the golden elaborations, and for the elaborations generated by two LLMs and two prompting strategies: (DP) Descriptive Prompt and (QUD) Questions Under Discussion.

3.3. Annotation Analysis

We measured inter-annotator agreement using Krippendorff's α (Krippendorff, 1980) and Fleiss' κ (Fleiss, 1971). Cohesion achieved moderate agreement ($\alpha = 0.61, \kappa = 0.61$), with unanimous judgments in 356 out of 500 instances (71.2%). In contrast, agreement for informativeness was fair ($\alpha = 0.24, \kappa = 0.08$), reflecting the more subjective nature of this criterion. Table 2 shows the distribution of informativeness labels in our dataset. Among the 226 instances judged as cohesive using annotator majority vote, agreement on informativeness was observed in only 38 cases (16.8%), with 26 labeled as *somewhat informative*, 5 as *uninformative*, and 7 as *informative*. This suggests that human annotators rarely agree on the informativeness of an elaboration, likely because this aspect is subjective and highly dependent on individual prior knowledge. Consequently, this resulted in significant data sparsity, which constrained our ability to conduct a fine-grained evaluation of informativeness. Nonetheless, these agreement levels provide a realistic reference point for our meta-evaluation.

The distribution of cohesion values across the final annotated dataset is presented in Table 3. The table indicates that elaborations generated by T5-Mistral were of lower quality relative to those produced by DeepSeek-R1-Qwen. Furthermore, the

results demonstrate that elaborations generated using the QUD approach yielded higher quality instances compared to the Descriptive Prompt approach. Surprisingly, five of the 'gold' instances from the ElabQUD corpus were in fact identified as incohesive by our annotators. Upon manual review, four of these five incohesive instances were confirmed to be indeed incohesive based on the context, suggesting that even reference corpora like ElabQUD contain instances that do not meet our cohesion criteria.[6] We further analyzed the 356 elaborations with unanimous cohesion judgments (166 cohesive and 190 incohesive). Manual inspection of the incohesive cases revealed several recurring issues: 103 elaborations contained text repetitions, 59 were logically inconsistent with the context, 27 included hallucinations, and 1 contained a misleading example.

4. Meta-evaluation of Automatic Metrics

In this section, we analyze how well existing strategies for evaluating elaboration quality correlate with human judgments, focusing on traditional reference-based metrics (Sec. 4.1) and (reference-free) LLM-based judgments (Sec. 4.2). For cohesion, we used a balanced subset of 462 instances with the same number of cohesive and incohesive instances. For informativeness, we used a subset of 146 instances that all annotators rated as cohesive, thus having three evaluations, where we also excluded instances that had extreme rating disagreements (i.e. being labeled as *uninformative* by one annotator and *informative* by another).

4.1. Reference-based Metrics

To the best of our knowledge, there are no automated metrics specifically tailored for evaluating elaborative simplifications. For our study, we selected three standard metrics similar to those utilized in earlier works (Srikanth and Li, 2021; Wu et al., 2023):

- BLEU (Papineni et al., 2002) measures the precision of n-grams in a candidate elaboration compared to a reference. We used the implementation available in the Evaluate library.[7]
- METEOR (Banerjee and Lavie, 2005) goes beyond exact word matches by incorporating stemming and synonymy, for a better measure of semantic equivalence. We used the implementation available in the Evaluate library.

[6] See Appendix D for a discussion of these instances.

[7] https://huggingface.co/docs/evaluate/index

	Cohesion AUC	**Informativeness** Spearman ρ
BERTscore F1	0.55 ± 0.12	0.26 ± 0.30
BLEU	0.55 ± 0.11	0.27 ± 0.32
Meteor	0.52 ± 0.12	0.19 ± 0.32

Table 4: Meta-evaluation of reference-based metrics.

- BERTScore (Zhang et al., 2020) leverages contextual embeddings from pre-trained language models to capture semantic similarity that goes beyond mere lexical overlap. In our study, we calculated the BERTScore F1 metric using *distilbert-base-uncased* as the contextual embedding model, relying on the implementation available in the Transformers library, chosen for better computational efficiency.[8]

These automatic metrics allow us to rank the elaborations from best to worst, and we assess how well these rankings agree with the human ratings. We used the gold elaborations from ElabQUD as reference texts and compared the automatic-metric scores with the collected human ratings. For cohesion, which is a binary feature in our dataset, we use the Area under the ROC Curve (AUC). For informativeness, we assess rank correlation between metric scores and the ordering induced by aggregated human ratings using Spearman's ρ. The results in Table 4 confirm that reference-based metrics perform poorly for elaboration evaluation. For cohesion, the AUC scores are close to the expected performance of random guessing (0.5). For informativeness, we see a weak (but statistically significant) positive correlation.[9]

4.2. LLM Judges

We experimented with 6 open-weight models of various sizes (falcon-3-7b, mistral-7b, gemma-3-4b-it, llama-3.1-8b, qwen3-next-80b, deepseek-chat-v3.1), and three closed-weight models (gpt5, gpt4o, gpt4o-mini). The LLMs were prompted with three configurations:

- **Full Guidelines**: prompts mirroring the comprehensive instructions given to human annotators;
- **Concise Prompt + CoT**: a condensed version of the guidelines (omitting examples) with a Chain-of-Thought (CoT) approach;
- **Full Guidelines + CoT**: the complete human guidelines combined with CoT.

[8] https://huggingface.co/docs/transformers/en/index

[9] p-value are (0.0014) for BERTscore, (0.0011) for BLEU, and (0.0197) for Meteor.

		Cohesion Acc	Informativeness Spearman ρ
Full Guidelines	gpt4o	**0.84** ± 0.06	0.53 ± 0.25
	qwen3-next-80b	0.82 ± 0.06	0.40 ± 0.29
	gpt5	0.78 ± 0.07	0.49 ± 0.25
	deepseek-chat-v3.1	0.75 ± 0.07	0.47 ± 0.28
	gpt4o-mini	0.64 ± 0.08	**0.54** ± 0.23
	falcon-3-7b	0.58 ± 0.09	0.48 ± 0.24
	gemma-3-4b-it	0.53 ± 0.09	0.27 ± 0.29
	llama-3.1-8b	0.50 ± 0.08	0.41 ± 0.28
	mistral-7b	0.48 ± 0.09	0.16 ± 0.35
Concise Prompt + CoT	gpt4o	**0.84** ± 0.06	**0.55** ± 0.24
	gpt5	0.81 ± 0.07	0.43 ± 0.27
	qwen3-next-80b	0.75 ± 0.08	0.46 ± 0.28
	deepseek-chat-v3.1	0.71 ± 0.08	0.48 ± 0.23
	gpt4o-mini	0.64 ± 0.08	0.52 ± 0.23
	falcon-3-7b	0.61 ± 0.09	0.48 ± 0.25
	gemma-3-4b-it	0.57 ± 0.08	0.46 ± 0.24
	llama-3.1-8b	0.54 ± 0.08	0.45 ± 0.27
	mistral-7b	0.50 ± 0.09	0.37 ± 0.27
Full Guidelines + CoT	gpt5	**0.86** ± 0.06	0.44 ± 0.26
	gpt4o	0.84 ± 0.06	**0.59** ± 0.22
	qwen3-next-80b	0.75 ± 0.07	0.46 ± 0.29
	deepseek-chat-v3.1	0.70 ± 0.08	0.41 ± 0.27
	gpt4o-mini	0.66 ± 0.08	0.50 ± 0.22
	falcon-3-7b	0.59 ± 0.08	0.48 ± 0.24
	llama-3.1-8b	0.56 ± 0.08	0.45 ± 0.26
	gemma-3-4b-it	0.53 ± 0.08	0.43 ± 0.26
	mistral-7b	0.47 ± 0.09	0.37 ± 0.28

Table 5: LLM evaluation results for cohesion and informativeness, with 95% confidence intervals.

The LLM-based evaluation does not rely on reference answers, and we separately prompt the models to assess cohesion and informativeness. For cohesion, the models generate a binary judgment. For informativeness, we tested two approaches: asking the models to use a 7-point scale (mimicking the combined human ratings) and asking for an informativeness degree between 0 and 100. For this analysis, we used the latter, as it performed better based on the average values of 3 runs.[10]

Table 5 shows that all three prompting strategies perform similarly. The Full Guidelines + CoT configuration produced the overall best results for evaluating cohesion, while Concise Prompt + CoT yielded the overall best results for informativeness. Comparing the different models, we can see a strong correlation between model size and the results for cohesion, with gpt5, gpt4o and qwen3-next-80b performing particularly well. In contrast, the smaller models are not capable of assessing cohesion to a meaningful extent, with performances around random guessing. For informativeness, the results show a moderate agreement with the human ratings, and the impact of model size is less clear (e.g. gpt4o-mini performing similarly to gpt4o).

[10]Appendix B compares the two evaluation approaches.

Error Type	Count
Text Repetitions	103
Logical Inconsistency	59
Hallucinations	27
Misleading examples	1

Table 6: Overview of errors in clear-cut "Incohesive" instances

Model	INC	REP	HAL
gpt4o-mini	5	5	0
gpt4o	2	1	0
gpt5	4	4	1

Table 7: Prevalence of different error types among the clear-cut cases of cohesion prediction: Logical Inconsistency (INC), Text Repetitions (REP) and Hallucinations (HAL).

4.3. Analysis

We further analyse the predictions of the GPT models, which performed best on our experiments.

4.3.1. Cohesion

To more effectively examine the false positives in the GPT models' cohesion predictions, we used a dedicated subset of 356 instances in which all annotators unanimously agreed on the cohesion label. Table 6 summarizes the error categories observed among the incohesive instances within this subset. Table 7 shows the frequency of each fault type among the elaborations that the GPT models failed to correctly identify as incohesive per error. While the overall number of false positives is small in these clear-cut cases, it is surprising that even gpt5 fails to detect some repetitions and logical inconsistencies, despite being explicitly prompted to regard such cases as incohesive.

We observed instances where even the largest models (gpt4o and gpt5) misclassified texts as cohesive despite the presence of text repetitions; Table 8 presents an example of this phenomenon. Furthermore, 10 cohesive instances that achieved the highest level of informativeness by the human annotators were unanimously identified as incohesive by all GPT models; an example of such a case is presented in Table 9. This indicates that despite being provided with the same guidelines as the human annotators, the LLMs did not fully understand the nature of the cohesion criterion.

CONTEXT: The reason why is complicated. They said its genes were too much like the genes of other giraffes. All plants and animals have genes. They play a big part in what animals and plants look and act like. Genes are passed down from parents.
ELABORATION: Genes are passed down from parents to children.

Table 8: Example of an incohesive instance containing text repetition misclassified as cohesive.

CONTEXT: Then about 6 million years ago another big change occurred. Big cats split into several different species. They became lions, tigers, jaguars and leopards. But there's a problem: What scientists find by looking at big cat DNA doesn't agree with what the fossils tell them. Scientists are hoping to figure out where big cats first appeared.
ELABORATION: This period was known as the Mesozoic Era. This was a time of great diversity and evolution.

Table 9: Example of a cohesive instance misclassified as incohesive.

4.3.2. Informativeness

Figure 2 plots gpt4o informativeness predictions against the combined human ratings. The model defaults to predict that elaborations are uninformative, with a clear cluster of predictions around a z-score of -1, and a smaller cluster around 0.5. Such values are also predicted for elaborations that humans judged as highly informative. In contrast, elaborations rated as the least informative by humans never receive the highest scores by the model. This shows that the model is more prone to underscoring informative elaborations, than to overscoring uninformative ones.

One challenge in this evaluation was the incompatibility between model predictions (0-100 scale) and the aggregated human judgments (discrete

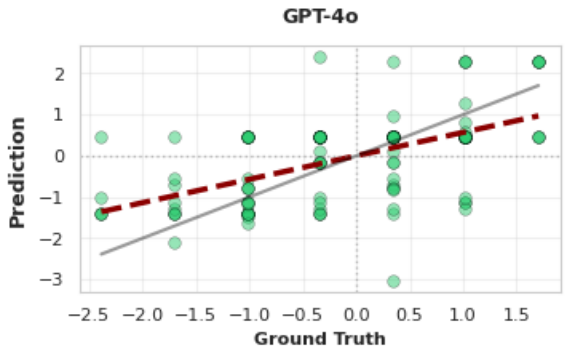


Figure 2: gpt4o Informativeness evaluation scatter plot.

CONTEXT: To reward his children, his father would bring books home. Usually, they were used books that cost just 5 cents. Billington is 85 years old. Even so, he does not plan to stop working anytime soon. He hopes to keep making it easier for people to use the library's huge collection.
ELABORATION: He feels a sense of accomplishment and satisfaction.

Table 10: Example of an instance underscored by all GPT models in informativeness.

3-9 scale). To enable a meaningful statistical comparison (e.g. for the scatter plot in Figure 2), we applied z-score standardization to both predicted and ground truth values.

Finally, Table 10 shows an example that was underscored by the GPT models. These instances tend to be short elaborations (around 10-15 words) that provide factual background information rather than direct explanations.

5. Conclusion

In this paper, we introduced a human-annotated dataset for the meta-evaluation of elaborative simplification evaluation methods. Our results show that standard reference-based metrics correlate weakly with human judgments, highlighting their clear limitations. LLM-as-a-judge approaches achieve higher correlation, provided that sufficiently large models are used. While their performance is impressive, especially in the light of the variability observed in human annotations, our qualitative analysis revealed that they remain an imperfect proxy for human judgment. Overall, the dataset provides a realistic reference for studying the strengths and limitations of automatic evaluation methods, especially as far as cohesion is concerned. Evaluating informativeness proved more challenging, as reflected by relatively low inter-annotator agreement. In future work, we will aim to address this by referring to a more clearly defined target audience in the definition of informativeness, and by splitting this criterion into more easily defined sub-criteria (e.g. How crucial is the information provided in this elaboration for understanding the text?, What proportion of the target audience would already be familiar with this information?).

Limitations

Due to budgetary constraints, we limited our scope by recruiting only three annotators and focusing our evaluation on elaborations generated from just two local LLMs. This restriction was required as the available funds were only sufficient to compensate

the participants for evaluating 500 instances each. Furthermore, given that our study focuses on elaborations generated by locally hosted LLMs, model selection was constrained to those compatible with the Nvidia RTX 4090 graphics card installed on the workstation used for model execution.

Ethics Statement

This study received a favourable ethical opinion from the School Research Ethics Committee. All participants filled consent forms containing instructions on how their data would be used. Annotators were compensated for their work with vouchers worth £125 each to annotate 500 instances, at a rate equivalent to the legal minimum wage. AI assistants were occasionally used in producing this work. AI tools were utilized for researching and identifying relevant research, code completion and optimization, and refining the writing through spell-checking and paraphrasing.

Lay Summary

To make complex writing easier to understand, a technique called "elaborative simplification" can be used. Instead of just swapping big words for small ones, this method adds extra context or explanations to help the reader follow along. However, it is difficult for researchers to measure if these added explanations are actually "good" or helpful.

To help with this, we created a new dataset that contains human ratings for explanations generated by AI. We asked people to judge these explanations based on two factors: cohesion (how well the explanations fit the context) and informativeness (how useful the added information actually is). We then tested whether other AI models could automatically grade these explanations as accurately as a human would.

Our findings show that while smaller AI models struggle to match human judgments, larger AI models can be quite effective at judging quality if they are given very specific instructions. We also found that "informativeness" is much harder to agree on than "cohesion," simply because what one person finds helpful, another might find unnecessary. This work provides a better roadmap for building AI tools that can explain complex ideas clearly and reliably.

6. Bibliographical References

Noof Abdullah Alfear, Dimitar Kazakov, and Hend Al-Khalifa. 2024. Meta-evaluation of sentence simplification metrics. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11229–11235, Torino, Italia. ELRA and ICCL.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of open language models.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. The (un)suitability of automatic evaluation metrics for text simplification. *Computational Linguistics*, 47(4):861–889.

Rohan Anil and Andrew M. Dai et al. 2023. Palm 2 technical report.

Sumit Asthana, Hannah Rashkin, Elizabeth Clark, Fantine Huot, and Mirella Lapata. 2024. Evaluating LLMs for targeted concept simplification for domain-specific texts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6208–6226, Miami, Florida, USA. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Abdullah Barayan, Jose Camacho-Collados, and Fernando Alva-Manchego. 2025. Analysing zero-shot readability-controlled sentence simplification. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6762–6781, Abu Dhabi, UAE. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023. Exploring the use of large language models for reference-free text quality evaluation: An empirical study. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 361–374, Nusa Dua, Bali. Association for Computational Linguistics.

Michael Chmielewski and Sarah C. Kucker. 2020. An mturk crisis? shifts in data quality and the impact on study results. *Social Psychological and Personality Science*, 11(4):464–473.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2024. Evaluating document simplification: On the importance of separately assessing simplicity and meaning preservation. In *Proceedings of the 3rd Workshop on Tools and Resources for People with REAding DIfficulties (READI) @ LREC-COLING 2024*, pages 1–14, Torino, Italia. ELRA and ICCL.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

J. L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li,

Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng,

Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning.

Yue Guo, Tal August, Gondy Leroy, Trevor Cohen, and Lucy Lu Wang. 2024. APPLS: Evaluating evaluation metrics for plain language summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9194–9211, Miami, Florida, USA. Association for Computational Linguistics.

David Heineman, Yao Dou, and Wei Xu. 2023. Thresh: A unified, customizable and deployable platform for fine-grained text evaluation. pages 336–345.

Freya Hewett, Hadi Asghari, and Manfred Stede. 2024. Elaborative simplification for German-language texts. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 29–39, Kyoto, Japan. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Klaus Krippendorff. 1980. Reliability. In *Content Analysis: An Introduction to Its Methodology*, pages 277–360.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Philippe Laban, Jesse Vig, Wojciech Kryscinski, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. 2023. SWiPE: A dataset for document-level simplification of Wikipedia pages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10674–10695, Toronto, Canada. Association for Computational Linguistics.

Zhen Li, Xiaohan Xu, Tao Shen, Can Xu, Jia-Chen Gu, Yuxuan Lai, Chongyang Tao, and Shuai Ma. 2024. Leveraging large language models for NLG evaluation: Advances and challenges. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16028–16045, Miami, Florida, USA. Association for Computational Linguistics.

Michael H. Long and Steven J. Ross. 1993. Modifications that preserve language and content. *EDRS*, pages 29–52.

Francesco Moramarco, Alex Papadopoulos Korfiatis, Mark Perera, Damir Juric, Jack Flann, Ehud Reiter, Anya Belz, and Aleksandar Savkov. 2022. Human evaluation and correlation with automatic metrics in consultation note generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1:*

Long Papers), pages 5739–5754, Dublin, Ireland. Association for Computational Linguistics.

Huyen Nguyen, Haihua Chen, Lavanya Pobbathi, and Junhua Ding. 2024. A comparative study of quality evaluation methods for text summarization.

OpenAI. 2024. GPT-4 technical report.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI*. Accessed: 2024-11-15.

Matthew Shardlow. 2014. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications(IJACSA), Special Issue on Natural Language Processing 2014*, 4(1).

Neha Srikanth and Junyi Jessy Li. 2021. Elaborative simplification: Content addition and explanation generation in text simplification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5123–5137, Online. Association for Computational Linguistics.

Mykola Trokhymovych, Indira Sen, and Martin Gerlach. 2024. An open multilingual system for scoring readability of Wikipedia. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6296–6311, Bangkok, Thailand. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ruiqi Wang, Jiyu Guo, Cuiyun Gao, Guodong Fan, Chun Yong Chong, and Xin Xia. 2025. Can LLMs replace human evaluators? an empirical study of LLM-as-a-judge in Software Engineering. *Proc. ACM Softw. Eng.*, 2(ISSTA).

BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Ja-

son Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Yating Wu, William Sheffield, Kyle Mahowald, and Junyi Jessy Li. 2023. Elaborative simplification as implicit questions under discussion. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5525–5537, Singapore. Association for Computational Linguistics.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Jeffrey T. Yarbro and Andrew M. Olney. 2021. Contextual definition generation. In *Proceedings of the Third International Workshop on Intelligent Textbooks 2021*, pages 74–83. CEUR-WS.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

A. Annotation guidelines

Task Description

To make a text easier to understand, adding more information can be helpful for readers. This added content is called an elaboration. There are various kinds of elaborations, such as explanations,

definitions, and other types of background information. For example: **Context** *"Photosynthesis is a foundational biological process that sustains nearly all ecosystems on the planet. Understanding it is vital for studying plant life as its efficiency impacts global atmospheric composition."* **Elaboration** *"This process, whereby plants convert light energy into chemical energy, also generates the oxygen we breathe."* In this task, you are asked to evaluate elaborations based on **two criteria**: *cohesion* and *informativeness*. They are defined as follows:

1. Cohesion:
A cohesive elaboration should:

- be relevant to the context;
- be free from errors and logical inconsistencies;
- not contain any misleading examples; and
- not simply repeat parts of the original text.

Cohesion is evaluated as **a binary property**, i.e. either an elaboration is *cohesive*, or it is *incohesive*. For example: **Context** "The government is split into two parties that often have different political beliefs."

Elaborations

- *"For example, the Labor Party and the Conservative Party."* **Cohesive.** The elaboration expands upon the original context with relevant and logically consistent information.
- *"This division often leads to lengthy debates and legislative gridlock, as each party attempts to push its own agenda."* **Cohesive.** The elaboration expands upon the original context with relevant and logically consistent information.
- *"The weather outside is very sunny."* **Incohesive.** The elaboration is irrelevant to the context provided.
- *"Therefore, the government is a single, unified entity with no internal disagreements."* **Incohesive.** The elaboration introduces logical inconsistencies for the context.
- *"The government is primarily focused on the production of chocolate and ice cream."* **Incohesive.** The elaboration introduces irrelevant information for the context.
- *"For example, the wedding party and the birthday party."* **Incohesive.** The elaboration introduces misleading examples.
- *"The government is split into two parties"* **Incohesive.** The elaboration repeats part of the original text.

2. Informativeness:
Informativeness evaluates the quality of the information provided in the elaboration based on the context. You will use a **scale from 1 to 3** based on how informative the elaboration is, as follows:

1. **Uninformative:** Nobody would benefit from the elaboration; it does not provide helpful information for understanding the context.
2. **Somewhat Informative:** Some people would benefit from the elaboration; it provides some basic information related to the context.
3. **Informative:** Most people would benefit from the elaboration; it goes beyond basic information by offering a more in-depth perspective or less obvious details.

For example: **Context** *"But not many countries support Obama's plan to fire missiles at Syria."*

Elaborations

1. Uninformative: *"Missiles can travel at speeds that exceed Mach 3."* The elaboration provides a general fact about missiles, but it does not help the reader understand the context of Obama's plan to fire missiles at Syria or the level of international support for it.
2. Somewhat Informative: *"Countries were concerned about the potential for civilian casualties."* The elaboration adds a relevant perspective by focusing on concerns from countries. It provides a basic reason for the opposition (civilian casualties), but lacks deeper insight into the political dynamics or the specific views of key stakeholders.
3. Informative: *"Allies of the Syrian government, might be drawn into the conflict, leading to a dangerous escalation."* The elaboration explains a key geopolitical risk, helping readers understand why countries opposed the plan. It adds depth by highlighting the potential for escalation, making it highly informative.

B. Additional Experimental Results

Table 11 shows the full classification metrics for cohesion evaluations, to complement the accuracy scores reported in the main paper. In particular, the table also shows the total number of instances predicted as cohesive and incohesive, as well as the precision, recall and F1 scores.

To evaluate LLM predictions of informativeness, we compared two approaches: asking the models to use a 7-point scale and asking for a degree between 0 and 100. Table 12 compares these two approaches.

Prompt & Models	Cohesive	Incohesive	Accuracy	Precision	Recall	F1
Full Guidelines						
gpt4o	231	231	0.844	0.844	0.844	0.844
qwen3-next-80b	192	270	0.825	0.891	0.740	0.809
gpt5	287	175	0.788	0.732	0.909	0.811
deepseek-chat-v3.1	120	342	0.751	0.983	0.511	0.672
gpt4o-mini	122	340	0.643	0.770	0.407	0.533
falcon-3-7b	351	111	0.580	0.552	0.841	0.667
gemma-3-4b-it	436	26	0.530	0.516	0.970	0.674
llama-3.1-8b	330	132	0.509	0.506	0.720	0.594
mistral-7b	111	351	0.485	0.469	0.228	0.307
Concise Prompt + CoT						
gpt4o	181	281	0.840	0.934	0.732	0.820
gpt5	250	212	0.816	0.792	0.857	0.823
qwen3-next-80b	130	332	0.751	0.946	0.532	0.681
deepseek-chat-v3.1	101	361	0.710	0.98	0.429	0.596
gpt4o-mini	107	355	0.640	0.804	0.372	0.509
falcon-3-7b	347	115	0.610	0.573	0.861	0.689
gemma-3-4b-it	382	80	0.578	0.547	0.905	0.682
llama-3.1-8b	342	120	0.543	0.529	0.784	0.632
mistral-7b	0	462	0.500	0.000	0.000	0.000
Full Guidelines + CoT						
gpt5	240	222	0.864	0.850	0.883	0.866
gpt4o	189	273	0.844	0.921	0.753	0.829
qwen3-next-80b	127	335	0.758	0.969	0.532	0.687
deepseek-chat-v3.1	96	366	0.704	0.990	0.411	0.581
gpt4o-mini	112	350	0.669	0.848	0.411	0.554
falcon-3-7b	359	103	0.593	0.560	0.87	0.681
llama-3.1-8b	338	124	0.565	0.544	0.797	0.647
gemma-3-4b-it	430	32	0.530	0.516	0.961	0.672
mistral-7b	255	207	0.476	0.478	0.528	0.502

Table 11: Full LLM Cohesion Evaluations

C. Experimental Details

The smallest LLMs were run locally on our test machines (falcon-3-7b[11], mistral-7b[12], gemma-3-4b-it[13], llama-3.1-8b[14]). For experiments with qwen3-next-80b and deepseek-chat-v3.1, we relied on the Openrouter[15] platform. Finally, gpt5, gpt4o and GPT-4o-mini were evaluated using the OpenAI API[16]. We used default hyperparameter values for the models.

[11] https://huggingface.co/tiiuae/Falcon3-7B-Instruct

[12] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.1

[13] https://huggingface.co/google/gemma-3-4b-it

[14] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[15] https://openrouter.ai/

[16] https://platform.openai.com

D. Discussion of Incohesive Golden Elaborations

In this section, we analyze the five 'gold' elaborations from the ElabQUD dataset that were rated as incohesive by our annotators out of the 100 total gold elaboration. As illustrated in Table 13 , with the exception of instance 3 in the table, which provided accurate information regarding the role of a jury, the remaining four instances are indeed incohesive. The faults are characterized by contextual irrelevance or logical inconsistencies. We attribute these errors to the extraction methodology used to identify elaborations within the original Newsela articles, which exclusively considers sentences immediately following a specified context. This approach occasionally captures textual fragments or the introductory phrases of unrelated sentences, leading to incoherent elaborations.

	7-point Scale Spearman ρ	**0 to 100 Scale** Spearman ρ
Full Guidelines		
gpt4o	0.50 ± 0.27	0.53 ± 0.25
qwen3-next-80b	0.36 ± 0.30	0.40 ± 0.29
gpt5	0.46 ± 0.25	0.49 ± 0.25
deepseek-chat-v3.1	0.46 ± 0.25	0.47 ± 0.28
gpt4o-mini	**0.52** ± 0.22	**0.54** ± 0.23
falcon-3-7b	0.42 ± 0.26	0.48 ± 0.24
gemma-3-4b-it	0.14 ± 0.33	0.27 ± 0.29
llama-3.1-8b	0.40 ± 0.27	0.41 ± 0.28
mistral-7b	0.25 ± 0.32	0.16 ± 0.35
Concise Prompt + CoT		
gpt5	0.49 ± 0.25	0.43 ± 0.27
gpt4o	0.48 ± 0.25	**0.55** ± 0.24
qwen3-next-80b	0.40 ± 0.27	0.46 ± 0.28
deepseek-chat-v3.1	0.38 ± 0.28	0.48 ± 0.23
gpt4o-mini	**0.50** ± 0.24	0.52 ± 0.23
falcon-3-7b	0.44 ± 0.28	0.48 ± 0.25
llama-3.1-8b	0.38 ± 0.27	0.45 ± 0.27
gemma-3-4b-it	0.12 ± 0.33	0.46 ± 0.24
mistral-7b	0.35 ± 0.33	0.37 ± 0.27
Full Guidelines + CoT		
gpt4o	0.42 ± 0.27	**0.59** ± 0.22
gpt5	0.47 ± 0.26	0.44 ± 0.26
qwen3-next-80b	0.39 ± 0.29	0.46 ± 0.29
deepseek-chat-v3.1	0.45 ± 0.26	0.41 ± 0.27
gpt4o-mini	**0.54** ± 0.22	0.50 ± 0.22
falcon-3-7b	0.48 ± 0.25	0.48 ± 0.24
gemma-3-4b-it	0.29 ± 0.31	0.43 ± 0.26
llama-3.1-8b	0.29 ± 0.31	0.45 ± 0.26
mistral-7b	0.43 ± 0.28	0.37 ± 0.28

Table 12: Informativeness Evaluation approaches comparison

	Context	Elaboration
1	Finn came up with a different explanation: Students cannot hide in the back of the classroom in smaller classes. They behave better and are more involved. He saw the change himself visiting classrooms in Buffalo, New York. Smaller, quieter classes may have their biggest effect on kids who do not pay attention and try to avoid looking the teacher in the eye. That's because they cannot hide.	But there was another question.
2	Steele is a scuba diver who has scooped up the animals from the seafloor since the 1970s. He sells the urchins to sushi restaurants. Steele heard about the sea getting more acidic. He saw right away what it could mean for his business and the ocean he loves. So Steele told Hofmann about the urchins.	The scientist started looking into his worries.
3	People listen to her, said Pastor Henry Logan, who has been working with her since August. "She does it one meal at a time." Last Monday, violent protests broke out again in Ferguson. A grand jury decided not to charge the officer who shot Brown. A grand jury is made up of a group of people.	They decide if a person should be charged with a crime.
4	He and his team named her Lucy after a song by the Beatles. The song played over and over the night her bones were found. Johanson is now the head of the Institute of Human Origins at Arizona State University. He spoke about how Lucy's discovery changed what scientists thought about early humans. He also discussed what he hopes to find next.	Newsela has adapted the answers given by Johanson.
5	They only made $27 more than what the city spent to put them on the streets. The Pasadena meters did not cost the city money. They were designed by college students. They were paid for by money that companies gave to the city, Huang said. City leaders say the money collected by the meters might help in finding people homes.	The charities have already proven to be very helpful.

Table 13: Incohesive Golden Elaborations

Readability Measures in Automatic Text Simplification: Is Simplification Quality a Coherent Construct?

Rémi Cardon[♦*], **A. Seza Doğruöz**[♥]
♦Computer Science and Engineering Department, UC3M, Spain
♥LT3, IDLab, Universiteit Gent, Belgium
rcardon@inf.uc3m.es, as.dogruoz@ugent.be

Abstract

Readability is a central concept in automatic text simplification (ATS), yet the two fields have largely developed in parallel, with limited cross-fertilization. While prior work has studied correlations between automatic evaluation metrics and human judgment in ATS, the correlations between these two aspects and readability measures have not received systematic attention. We address this gap by investigating to what extent readability measures align with both human judgment and automatic metrics in ATS. Using two English datasets annotated with human judgments (SimplicityDA at the sentence level and D-Wikipedia at the document level), we compute 1,066 linguistic features (covering lexical diversity, lexical sophistication, syntactic sophistication, and cohesion) and eight traditional readability formulas, and correlate them against human scores and standard ATS metrics (BLEU, SARI, BERTScore, LENS, D-SARI). Our results show that readability measures correlate poorly with both human judgment and automatic metrics across both levels. The meaning preservation criterion consistently yields the highest correlation values, while simplicity and fluency criteria remain low. We also find systematic differences between sentence-level and document-level simplification in terms of which features are most informative: type-token ratio features are predictive at the sentence level but not at the document level, while corpus-frequency features show the opposite pattern. These findings point to a broader issue: ATS lacks a shared theoretical construct for simplification quality, and the three main approaches to its assessment (human judgment, readability measures, and automatic metrics) do not consistently converge.

Keywords: Automatic Text Simplification, Readability, Evaluation

1. Introduction

The accessibility of written information is an important question: outside natural language processing (NLP), domains like medicine (Gu et al., 2024) or business (Huong Dau et al., 2024) have been studying the readability of the documents they produce (e.g., medical reports or information for patients, business reports for shareholders). Usually, those studies are performed using traditional readability formulas, like the Flesch Reading Ease (Flesch, 1948) or Dale-Chall (Dale and Chall, 1948) formulas mainly developed for written English texts. Recently, the reliability of these formulas has been questioned (Alzaid et al., 2024). In NLP, automatic text simplification (ATS) aims at transforming texts to make them more accessible, while preserving their meaning (Saggion, 2017).

Since ATS aims at making texts more accessible, readability is a natural way to frame its goal. However, it is not clear to what extent readability measures and judgments on ATS systems (e.g., human judgment or automatic metrics) correlate with each other. In this paper, we investigate the position of readability measures in the ATS landscape. While readability is regularly mentioned in ATS studies, ATS and automatic readability assessment (ARA) have largely developed in parallel, with limited cross-fertilization (Vajjala, 2022).

While there have been studies on the correlations between ATS evaluation metrics and human judgment (Alva-Manchego et al., 2021; Cripwell et al., 2024), the correlations between these two aspects and commonly available readability measures have not been studied. Our research question is: *to what extent do readability measures correlate with (a) human judgment and (b) automatic evaluation metrics in ATS?*

We are interested in this research question because readability is used differently across ATS research. Some studies incorporate a formula (e.g., FKGL) directly into a training loss (Flores et al., 2023), others include CEFR (Common European Framework of Reference for Languages[1]) levels in LLM prompts (Imperial and Tayyar Madabushi, 2023; Maddela and Alva-Manchego, 2025), and others compute hundreds of linguistic features to characterize a corpus (Battisti et al., 2020; Vajjala and Lučić, 2018), with no shared rationale for why any of these choices should reflect simplification quality.

We argue that this inconsistency is symptomatic of a broader issue: ATS lacks a clear *construct* for simplification quality. The term construct (bor-

*Research partially done while employed at UGent.

[1] https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions

Proceedings of the Joint Workshop on Readability and Text Simplification (READIxTSAR) @ LREC 2026, pages 210–226
11 May 2026.

rowed from psychometrics) refers to the theoretical concept that an evaluation instrument is supposed to measure. In ATS research, there is a need to define what "simplification quality" refers to before attempting to measure it. Lexical complexity, syntactic structure, meaning preservation and fluency are all dimensions that readability features and ATS metrics capture separately, but there is no shared definition of what their combination is supposed to measure. It is hard to tell whether two systems that score differently on BLEU and SARI metrics are actually different in quality, or they are just optimizing for different aspects of the same vague objective. ATS research relies on three main approaches to evaluate simplification quality: human judgment, readability measures, and automatic metrics. We empirically test whether these converge, as a necessary condition for treating them as proxies for the same underlying construct. Any divergence would not render them useless but would reinforce the need for a clearer construct definition.

More specifically, this paper makes the following contributions: (i) a correlation study between readability measures and human judgment on two English datasets (sentence-level and document-level); (ii) a correlation study between readability measures and standard ATS evaluation metrics; (iii) a discussion of implications for construct definition in ATS.

2. Related Work

In this section, we provide some background information about readability and ATS research (Sections 2.1, 2.2) before discussing how the two fields interact with each other (Section 2.3).

2.1. Readability

Readability research dates back to the 1920s, initially motivated by the need to assess the suitability of texts for school-age readers (Lively and Pressey, 1923). This first method relied on a list of word frequencies (Thorndike, 1921), based on the assumption that texts made of frequent words are more readable. For a more detailed historical overview, see François (2015). We briefly summarize the key periods below.

The early period of readability research consisted of identifying predictors and tuning coefficient weights out of corpus-based observations and annotations by humans. The most famous readability formulas for English are Flesch Reading Ease (Flesch, 1948, FRE) and Flesch-Kincaid Grade Level (Kincaid et al., 1975, FKGL), which rely on word count and number of syllables per word.

Early NLP-based approaches (1990s–2000s) relied on regression, latent semantic analysis and language modeling (Daoust et al., 1996; Foltz et al., 1998; Si and Callan, 2001). In the 2020s, ARA (automatic readability assessment) has developed into a lively line of research (Vajjala, 2022). ARA has been explored with distributional text representations and with linguistic features. The distributional text representations follow the advancements of research in machine learning, notably with the development of transformers (Vaswani et al., 2017). Regarding linguistic features, the way to select and leverage them is still an open question. Nonetheless, research on this question is facilitated by the appearance of tools that can be used to compute an increasingly high number of features, for example for English (Kyle et al., 2021, 2018; Lu, 2010; Crossley et al., 2019) or French (Wilkens et al., 2022). These tools produce raw analyses of texts with hundreds of features but there are no recommendations about how to select and use them. This gap has fueled new research aimed at combining linguistic feature vectors with distributional representations to improve automatic readability assessment, based on the hypothesis that the two types of information are complementary (Deutsch et al., 2020; Lee et al., 2021; Wilkens et al., 2024).

The readability features depend heavily on the language that is under study, and the aforementioned tools rely on language-dependent resources such as reference corpora, vocabulary lists, or pre-trained models (e.g., for POS-tagging or syntactic analysis). So far, most studies on readability focus on English, due to the availability of tools and resources.

2.2. Automatic Text Simplification

In this section, we briefly describe ATS to provide a background for the discussion of how it integrates considerations about readability, as covered in Section 2.3.

Methods ATS has traditionally been performed at the sentence level (Saggion, 2017) and continues to be an active area of research (Kew et al., 2023). In the early works, the goal was to make sentences simpler to handle as an input for other NLP systems such as syntactic parsers (Chandrasekar et al., 1996). It was only later explored as a means of simplifying texts to make them easier to understand by humans (Carroll et al., 1999). These initial computational methods were rule-based and targeted only specific operations in a sentence (Cardon and Bibal, 2023) (e.g., removing appositive clauses, changing the voice of a sentence from passive to active). The recent developments of generative models have shifted ATS research

towards document-level simplification (Sun et al., 2021), notably with multi-agent architectures (Mo and Hu, 2024; Fang et al., 2025).

Evaluation. Evaluation of ATS is an open question. Traditional readability formulas, mostly FKGL or adaptations of FRE for other languages, are often reported (Engelmann et al., 2024; Flores et al., 2023; Aluisio et al., 2010; Paula and Camilo-Junior, 2024; Štajner and Saggion, 2013). However, they correlate poorly with human judgment on ATS (Tanprasert and Kauchak, 2021; Alva-Manchego et al., 2021). The most common automatic metrics are BLEU (Papineni et al., 2002), SARI (Xu et al., 2016), BERTScore (Zhang et al., 2020) and LENS (**?**). D-SARI (Sun et al., 2021) extends SARI to the document level. BLEU and BERTScore compare the text output to one or more references, while (D-)SARI adds the input into the computation. Although BLEU is often interpreted as an indicator of meaning preservation, SARI of simplicity, and BERTScore of meaning preservation and fluency, prior work has shown that these metrics correlate poorly and inconsistently with human judgment (Alva-Manchego et al., 2021; Sulem et al., 2018a).

These three criteria (fluency, simplicity and meaning preservation) are also used by human annotators to evaluate sentence simplification, typically on 5-point Likert scales. For document-level simplification, methods for human evaluation are not stabilized yet. Cripwell et al. (2024) use the same criteria with binary questions instead of Likert scales. Sun et al. (2021) ask human judges to evaluate "overall simplicity" that they define as simplicity with other quality criteria (e.g., ease of reading and meaning preservation). Vásquez-Rodríguez et al. (2023) ask human judges to evaluate textual coherence and Agrawal and Carpuat (2024) evaluate meaning preservation by studying human accuracy on reading comprehension questions about the simplified text.

2.3. Readability and Text Simplification

In most research where readability and ATS interact, readability is leveraged through linguistic features to give information about the datasets (Battisti et al., 2020; Vajjala and Lučić, 2018; Yaneva et al., 2016; Štajner and Saggion, 2013; Dell'Orletta et al., 2011; Aluisio et al., 2010). Other studies rely on linguistic features for data selection instead (Jingshen et al., 2024). Jingshen et al. (2024) leverage readability features, in conjunction with similarity measures, to mine sentence pairs and produce a parallel corpus for Chinese idiom simplification. De Martino (2023) investigates the link between eye-tracking data and readability features on Italian data. Although preliminary, this study suggests that eye-tracking is a promising method for assessing the cognitive processing cost of simplified texts, and could complement annotation-based measures of simplification quality.

Some simplification studies use readability features or metrics in their evaluation protocol as well. Scholz and Wenzel (2025) evaluate 18 readability features (i.e., syntactic, POS-based, semantic and fluency features) for English and German text simplification. They find that some measures (e.g., semantic and fluency features) transfer across languages, while the behavior of statistical, POS-based and syntactic metrics appears to be strongly language-dependent. Paula and Camilo-Junior (2024) use a Portuguese adaptation of FRE as an evaluation metric for ATS. Engelmann et al. (2024) use the FRE and Dale-Chall formulas to perform pairwise comparisons to rank simplifications. They compare them to human judgments and GPT 3.5 and find that Dale-Chall has the highest correlation to human judgment, above GPT 3.5, while FRE obtains the lowest correlations.

Readability can also be incorporated in ATS methods. Flores et al. (2023) use a bounded FKGL as a component of their loss in a neural model for text simplification. Maddela and Alva-Manchego (2025) and Imperial and Tayyar Madabushi (2023) prompt LLMs for document-level simplification by including CEFR levels in the prompt. Using CEFR as a proxy for readability was introduced with the release of the CEFR-SP dataset (Arase et al., 2022), a corpus of 17,000 English sentences annotated according to CEFR levels.

Readability features have also been used at a more granular level. Lexical complexity features, in particular, have been leveraged for lexical simplification (North et al., 2025). Hazim et al. (2022) introduce a system that highlights complex words in a text editor to help humans manually simplify texts. Maddela and Xu (2018) use lexical features to rank candidates for substitution in a neural lexical simplification system. Grigonyte et al. (2014) rely on lexical and morphological features to perform complex word identification.

It is worth noting that readability and ATS studies address related but distinct objectives. Readability research primarily targets lexico-syntactic properties of texts and their effect on comprehension by humans regardless of the source text. On the other hand, ATS transforms a source text into a simplified version. Its evaluation typically involves comparing the output to the input (or to a reference). These different perspectives motivate our research question: whether readability measures are aligned with how simplification quality is assessed, both by humans and by automatic metrics.

Overall, we observe that various readability measurement approaches (e.g., linguistic features, for-

mulas, eye-tracking, CEFR levels) are also explored in ATS research. Table 1 provides a structured overview of ATS research and illustrates the diversity of practices and the lack of a shared theoretical framework.

The two approaches most widely present in ATS are traditional formulas (e.g., FRE and FKGL, used as evaluation metrics) and readability features (for providing information about datasets), both of which we examine in this study.

3. Readability Measures and ATS Metrics

Throughout this paper, we use *features* to refer to linguistic features, *formulas* to refer to traditional readability formulas (e.g., FRE, FKGL), *measures* as an umbrella term for features and formulas, and *metrics* for automatic ATS evaluation metrics (e.g., BLEU, SARI).

3.1. Data

We focus on English because it is easier to find (i) human-annotated simplification datasets and (ii) feature-extraction tools for our study of correlations between readability measures and ATS evaluation criteria.

To study how readability measures correlate with the evaluation protocols in ATS, we rely on two datasets: at the sentence level (Alva-Manchego et al., 2021) and at the document level (Maddela and Alva-Manchego, 2025). They are both labeled with human judgment. Both works studied the link between automatic metrics and human judgment. Building on their findings, we explore (a) the link between readability measures and human judgment, and (b) the link between readability measures and automatic metrics. Below is a description of the datasets.

SimplicityDA. For the sentence-level study, we use Simplicity-DA (Alva-Manchego et al., 2021)[2]. It consists of 600 sentence-level ATS system outputs in English, all annotated by 15 crowdworkers based on three criteria (fluency, simplicity and meaning preservation) on a 0-100 scale. Besides the human judgments, the dataset also includes automatic scores (e.g., BLEU, SARI, BERTScore and SAMSA (Sulem et al., 2018b)) for each sentence.

D-Wikipedia. For the document-level study, we use D-Wikipedia (Sun et al., 2021). D-Wikipedia is a corpus of aligned paragraph pairs extracted from the English Wikipedia for the complex side and Simple English Wikipedia for the simple side. Maddela and Alva-Manchego (2025) released a subset of 100 paragraph pairs from D-Wikipedia, each with 4 simplified versions produced by automatic systems, resulting in 500 paragraph pairs. These 500 pairs were rated by three human judges on a 5-point Likert scale on three criteria (fluency, simplicity and meaning preservation). We compute the automatic metrics values (e.g., BLEU, SARI, D-SARI, BERTScore and LENS) with the code provided with the dataset[3].

3.2. Readability Measures

Readability Features. As discussed in Section 2, readability assessment is mostly explored with two types of text representations (distributional embeddings and textual features). As distributional embeddings are already leveraged for simplification methods (Kew et al., 2023) and evaluation (Zhang et al., 2020), we focus on textual features. To compute these features, we use four tools that implement a total of 1,066 readability-related features for English[4].

TAALED (Kyle et al., 2021)[5] computes 38 features related to lexical diversity, such as different type-token ratios or MTLD (Measures of Textual Lexical Diversity).

TAALES (Kyle et al., 2018)[6] computes 484 features related to lexical sophistication, i.e., the degree to which the vocabulary used in a text is advanced, infrequent, or complex relative to a reference population. Many of these features are variations of word frequency (computed on various corpora such as BNC (Consortium, 2007, The British National Corpus) and COCA (Davies, 2008, The Corpus of Contemporary American English)). Other features are related to lexical neighborhood (i.e., the number of words that are orthographically or phonologically similar to a given word, a measure linked to word recognition difficulty), age of acquisition (i.e., the estimated age at which a word is typically learned by native speakers, a correlate of word difficulty), psycholinguistic norms (e.g., concreteness, imageability, meaningfulness).

TAASSC (Lu, 2010)[7] computes 376 features related to syntactic sophistication. These features rely on grammatical dependency analysis and part-of-speech tagging. Some examples of these fea-

[2] https://github.com/feralvam/metaeval-simplification

[3] https://github.com/cardiffnlp/document-simplification

[4] The complete list of features and their formulas is available in the documentation of each tool.

[5] https://www.linguisticanalysistools.org/taaled.html

[6] https://www.linguisticanalysistools.org/taales.html

[7] https://www.linguisticanalysistools.org/taassc.html

Usage	Reference	Language(s)	Measure type
Corpus description	Aluisio et al. (2010)	Portuguese	Features / Formulas
	Dell'Orletta et al. (2011)	Italian	Features
	Štajner and Saggion (2013)	Spanish	Features / Formulas
	Yaneva et al. (2016)	English	Features
	Vajjala and Lučić (2018)	English	Features
	Battisti et al. (2020)	German	Features
	De Martino (2023)	Italian	Features + eye-tracking
Corpus construction	Jingshen et al. (2024)	Chinese	Features
Evaluation	Engelmann et al. (2024)	English	Formulas (FRE, Dale-Chall)
	Scholz and Wenzel (2025)	German	Features
Method	Grigonyte et al. (2014)	Swedish	Lexical features (CWI)
	Maddela and Xu (2018)	English	Lexical features
	North et al. (2025)	English	Lexical features
	Hazim et al. (2022)	Arabic	Lexical features
	Flores et al. (2023)	English	Formula (FKGL in loss)
	Paula and Camilo-Junior (2024)	Portuguese (Braz.)	Formula (FRE as target)
	Imperial and Tayyar Madabushi (2023)	English	CEFR (in prompt)
	Maddela and Alva-Manchego (2025)	English	CEFR (in prompt)
	Barayan et al. (2025)	English	CEFR (in prompt)

Table 1: Overview of works at the intersection of readability and ATS, grouped by how readability measures are used. CWI = Complex Word Identification.

tures are conjunctions per clause, verbal modifiers per noun phrase, frequency of constructions compared to references coming from different corpora (e.g., BNC, COCA and others) or more traditional ones (e.g., average sentence length).

TAACO (Crossley et al., 2019)[8] computes 168 features related to cohesion. Some examples include semantic similarity between word2vec (Mikolov et al., 2013) embeddings of adjacent sentences, or token overlap between adjacent sentences or paragraphs.

These four tools cover complementary dimensions of text complexity. Specifically, TAALED and TAALES focus on lexical aspects (diversity and sophistication, respectively), TAASSC targets syntactic structure, and TAACO captures discourse-level cohesion. Together, they cover the main dimensions associated with readability in the literature, providing a broad and principled basis for our correlation study.

Readability Formulas. We also compute the following set of traditional readability formulas for English, using the `textstat` Python library: Flesch Reading Ease (Flesch, 1948), Dale-Chall (Dale and Chall, 1948), Gunning-Fog (Gunning, 1952), Linsear Write (O'hayre, 1966), ARI (Smith and Senter, 1967), SMOG (Mc Laughlin, 1969), Flesch-Kincaid Grade Level (Kincaid et al., 1975), and Coleman-Liau (Coleman and Liau, 1975).

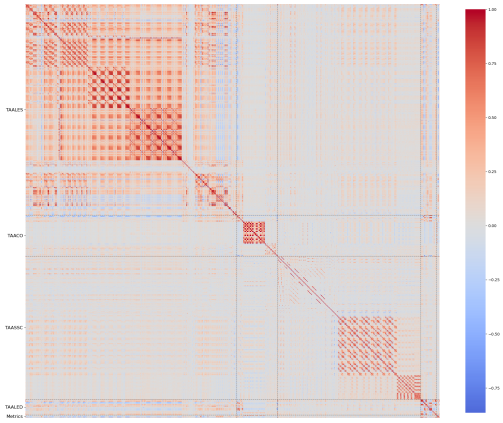
[8] https://www.linguisticanalysistools.org/taaco.html

4. Experiments

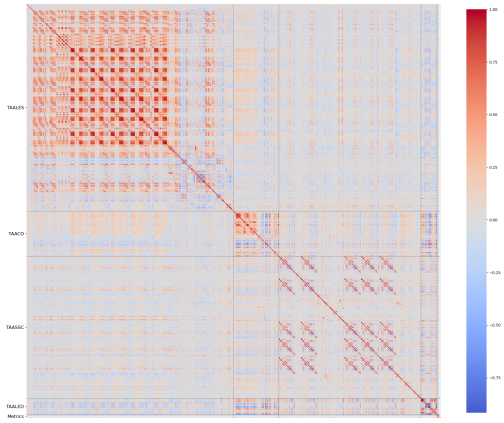
4.1. Readability Measures

First, we compute the correlations among the readability measures themselves (i.e., the 1,066 linguistic features from the four tools and the eight traditional readability formulas described in Section 3.2). Figures 1a and 1c show the correlation matrices computed on the SimplicityDA dataset (at the sentence level), respectively on the difference between the simplified and original sentences, and on the simplifications. Figures 1b and 1d show the correlation matrices computed on the D-Wikipedia dataset, respectively on the difference between the simplified and original sentences, and on the simplifications. Our findings include: (i) the measures mostly correlate with other measures of the same type, (ii) measures computed at the document level show higher absolute values and (iii) measures computed on the difference between original texts and simplifications exhibit lower absolute values.

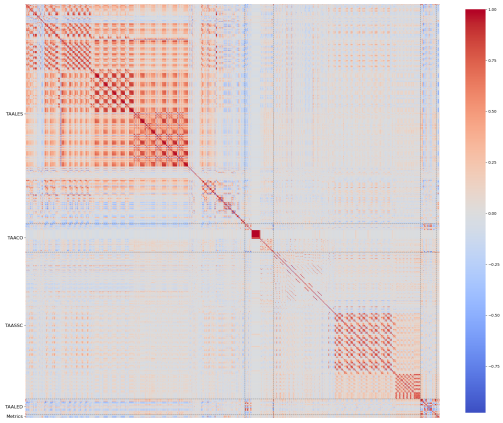
These observations are visible in Figure 1: in all four heatmaps, the block structure along the diagonal confirms that intra-group correlations dominate. The higher saturation in the document-level heatmaps (Figures 1b and 1d) visually reflects observation (ii). In the delta heatmaps (Figures 1a and 1b), the overall lighter coloration reflects observation (iii): subtracting original from simplified values reduces systematic co-variation between measures. Consequently, the few correlation clusters that remain visible in these heatmaps correspond to measures that respond differently to the



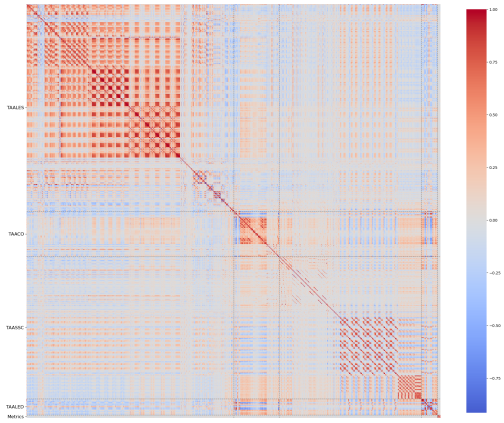
(a) Difference between original and simple, SimplicityDA (sentence-level).



(b) Difference between original and simple, D-Wikipedia (document-level).



(c) Simple side of SimplicityDA (sentence-level).



(d) Simple side of D-Wikipedia (document-level).

Figure 1: Pearson correlation matrices of readability measures and metrics. Dashed lines indicate the boundaries of feature groups (from top to bottom, and the same from left to right: TAALES, TAACO, TAASSC, TAALED, and Metrics).

Dataset	Mode	Criterion	Top $\|r\|$	10th $\|r\|$
SimplicityDA	simp	simplicity	.199	.159
	simp	fluency	.262	.195
	simp	meaning	.295	.231
	delta	simplicity	.215	.150
	delta	fluency	.300	.211
	delta	meaning	.426	.350
D-Wikipedia	simp	simplicity	.205	.091
	simp	fluency	.103	.096
	simp	meaning	.433	.192
	delta	simplicity	.298	.234
	delta	fluency	.291	n=4
	delta	meaning	.342	.211

(a) Human judgment correlations.

Dataset	Mode	Metric	Top $\|r\|$	10th $\|r\|$
SimplicityDA	simp	BERTScore	.339	.273
	simp	BLEU	.352	.290
	simp	SAMSA	.309	.253
	simp	SARI	.378	.322
	delta	BERTScore	.518	.413
	delta	BLEU	.403	.282
	delta	SAMSA	.199	.184
	delta	SARI	.499	.410
D-Wikipedia	simp	D-SARI	.220	.129
	simp	LENS	.347	.185
	simp	BERTScore	.143	.090
	delta	D-SARI	.305	.251
	delta	LENS	.279	.217
	delta	BERTScore	.219	.179

(b) Automatic metrics correlations.

Table 2: Top absolute $|r|$ values among significant Pearson correlations between readability measures and evaluation criteria. "simp" = simple version only; "delta" = difference simple−original; 10th $|r|$ = lowest of the top 10 values (n=4: fewer than 10 significant correlations found). Full tables per feature in Appendix A.1.

simplification operation itself. Therefore, they are considered to be the most informative ones for our study.

4.2. Measures and Human Judgment

To compare readability measures (the features with the four readability tools, and the readability formulas) and human judgment, we compute them all on both datasets: SimplicityDA for the sentence-level (100 original sentences and 600 simplifications including 100 human-written ones) and D-Wikipedia for the document-level (100 original paragraphs and 500 simplifications including 100 human-written ones). For each dataset, we compute the measures on both sides (original and simplified) separately. We compute the correlations with human judgment in two ways: (i) on the measures obtained on the simplified versions only, and (ii) on the difference between the measures obtained on the original texts and the ones obtained on the simplified versions. The first case focuses on simplicity, and the second on simplification by including a comparison with the original text.

For both datasets, we report the correlations on the three criteria (simplicity, fluency and meaning preservation) for human judgment.

4.3. Measures and Automatic Metrics

Metrics like BLEU and SARI measure string overlap rather than linguistic complexity. However, testing whether readability features correlate with them is also informative when researchers use these metrics as proxies for measuring simplification quality. Readability features can also be used for the same purpose; a low correlation would confirm that they capture fundamentally different aspects of the output.

To study the correlations between readability measures and automatic simplification metrics, we proceed in the same way as for the correlations between readability measures and human judgment. We report scores on the following automatic metrics: BLEU, SARI, BERTScore for SimplicityDA (sentence-level), and D-SARI, BERTScore, and LENS for D-Wikipedia (document-level).

Regarding the metrics that require references (BLEU, SARI), we use all the references that are available in Simplicity-DA (i.e., for each original sentence: 10 references from ASSET (Alva-Manchego et al., 2020), 1 from TurkCorpus (Xu et al., 2016) and 1 from HSplit (Sulem et al., 2018a)). For D-Wikipedia, we use the single reference simplification that is provided for each original text.

5. Results

5.1. Measures and Human Judgment

We report the top significant correlations between readability measures and human judgment in Table 2a (full tables per dataset and criterion in Appendix A.1).

For SimplicityDA, the highest absolute coefficient values are obtained with the meaning criterion computed on delta, with the top 10 ranging from -0.43 to -0.35. All of the other criteria have top absolute coefficient values between 0.15 and 0.30. In that regard, readability measures (features and formulas) and human judgment on simplification quality do not correlate well at the sentence level. We also observe that all absolute values are higher when computed on the delta rather than on simplifications only. As the human judges were asked to rate simplification instead of simplicity, the difference between simplicity and simplification has an effect on both humans and measures (while the coefficient values are low).

These results call for caution in using readability features as standalone indicators of simplification quality. They can help characterize corpora or contribute to composite metrics, but they should not be read as direct proxies for what human judges find simple.

Regarding the D-Wikipedia dataset, the observations are similar. Meaning exhibits the highest coefficient values, although with a higher discrepancy between the top 1 and 10 values (.433 vs .192 for simp-meaning and .342 vs .211 for delta-meaning). We found only 4 significant correlations for delta-fluency, suggesting limited correlation at best (the highest values being .291 for delta and .103 for simp). As for SimplicityDA, the values are generally higher for delta than for simp.

The observation that only 4 significant correlations were found for delta-fluency on D-Wikipedia is notable. It may reflect either that grammaticality is not captured by the features we use, or it is harder to detect it at the document level due to greater variability in the simplifications.

Regarding simplicity, the values are low for both datasets. The most correlated set of observations is delta-simplicity with top 10 absolute values ranging from .234 to .298.

Not many features are found in more than one set of highest correlating values. For SimplicityDA, we observe several kinds of type/token ratio (TTR). Root TTR and log TTR are the most recurrent, appearing in 3 and 4 sets of observations out of 6, respectively. For D-Wikipedia, we see that the word count appears in 4 sets of observations, and corpus-based metrics (especially calibrated on COCA but also on the BNC) appear in 5 out of 6 sets.

5.2. Measures and Automatic Metrics

We report the correlations between readability measures and automatic metrics in Table 2b (full tables in Appendix A.1).

For SimplicityDA, BERTScore has the highest correlation values, especially when the features are computed on delta (with a top 10 ranging from .413 to .518). SAMSA exhibits the lowest correlation values and is the only metric to correlate better when the features are computed on the simple texts only. SAMSA measures structural simplification (sentence splitting), which correlates with higher token counts on the simple side rather than with the *reduction* of linguistic features that other metrics reward. It is an example of how different metrics operationalize different sub-tasks of simplification, reinforcing the argument for a clearer construct definition.

Regarding D-Wikipedia, the correlations are generally lower. BERTScore has the lowest correlation values (from 0.09 to 0.219 across both computation modes), while LENS exhibits a slightly higher level of correlation than D-SARI.

Regarding the features themselves, COCA-based features are present in all criteria with D-Wikipedia, while they are only present for delta-SAMSA with SimplicityDA.

TTR measures are present in 5 out of 8 sets of observations for SimplicityDA, and are completely absent for D-Wikipedia. These observations suggest that sentence simplification and document simplification evaluation do not involve the same phenomena. It is worth noting that TTR correlates better with sentence-level simplification than with document-level simplification, even though TTR is commonly used as a rough measure of text complexity elsewhere.

5.3. Cross-level Observations

Comparing the two datasets reveals differences between sentence-level and document-level simplification that are not just a matter of scale. For human judgment, the overall pattern holds at both levels. The meaning criterion yields the highest correlation values, while the simplicity and fluency criteria remain low. This suggests that readability features are more sensitive to semantic distortion than to structural or fluency changes, regardless of granularity.

The divergence is more visible in which features matter. TTR measures are among the top correlations for SimplicityDA (present in 5 out of 8 sets), but completely absent for D-Wikipedia. This is consistent with a known limitation of TTR. It is sensitive to text length, since longer texts tend to accumulate repetitions that drive TTR down, irrespective of lexical richness. At the document level (where texts are longer and more varied), TTR loses its discriminative power. Conversely, frequency-based features calibrated on COCA appear across all criteria for D-Wikipedia, but only marginally for SimplicityDA. Corpus frequency measures may capture the lexical choices that distinguish document-level simplification outputs better.

These observations suggest that the readability features for sentence-level and document-level evaluation are not the same. Feature selection should be adapted to the granularity of the task rather than applying it uniformly.

6. Conclusion

This paper investigated how readability measures correlate with human judgment and automatic metrics in ATS. The question is relevant because the field already uses readability measures in evaluation protocols and training objectives, often without theoretical grounding, and because prior work has documented inconsistent correlations between human judgment and automatic metrics. We acknowledge that our findings do not go towards dissipating the uncertainty that the field has been experiencing. That said, our findings point to a lack of a well-defined construct in the ATS ecosystem. In the case of ATS, the question is whether "simplification quality" is a coherent, measurable construct, or whether it bundles together with other criteria (e.g., fluency, lexical simplicity, structural changes, meaning preservation) that different metrics and features capture in different ways. Our results show that readability measures, automatic metrics and human judgments do not consistently converge. When a system performs lexical simplification, lexical sophistication features will improve and FKGL may also improve, but SARI may penalize the output if the reference chose a different synonym. In practice, a simplified text results from dozens of such operations pulling in different directions, and each measure captures only one facet of the result. Without a shared construct, none of these scores can be read as unambiguous evidence of quality. This suggests that the field has not yet settled on a shared construct for simplification quality, and working towards such a definition would benefit both evaluation design and system development.

7. Limitations

The main limitations of our work concern the quantity and quality of data. We used the only data with human judgment that were available to us in English. Both datasets have limitations. SimplicityDA covers outputs from a limited number of systems, which may not reflect the full diversity of simplification approaches. D-Wikipedia is based on a

small annotated subset (100 paragraph pairs), with only three annotators per pair. These factors may contribute to the low correlations we observe independently of any conceptual mismatch between constructs. These findings may therefore vary on higher-quality or more recent corpora, on other languages, and with other human annotators. While this impairs the generalizability of our study, it reinforces our point that the field should, as a community, focus more on clearly defining the task and producing better-controlled evaluation data.

A further limitation is the univariate nature of our analysis. We compute each feature's correlation independently, which does not capture interactions between features. Methods such as principal component analysis or multivariate regression could reveal richer structure in the relationship between readability features and evaluation outcomes.

8. Acknowledgements

We would like to thank the reviewers for their valuable comments. Rémi Cardon is partially funded by grant PID2023-148577OB-C21 (Human-Centered AI: User-Driven Adapted Language Models-HUMAN_AI) by MICIU/AEI/10.13039/501100011033 and by FEDER/UE.

9. Plain Language Summary

Making written information easier to read is important, especially when using computers to rewrite complex texts. This process is called automatic text simplification (ATS). However, experts do not always agree on how to measure if a text has really become easier to read. There are three main ways to check: asking people to judge the text, using computer formulas that estimate readability, and using automatic scores that compare the original and simplified texts.

In our study, we wanted to see if these three ways of measuring simplification quality actually agree with each other. We used two collections of English texts that had already been rated by people for how simple, fluent, and meaningful they were. We also used computer programs to calculate over a thousand different features of the texts, as well as eight common readability formulas.

We found that readability formulas and features do not match well with what people think, or with the automatic scores. The only area where there was some agreement was in how well the meaning of the text was preserved. We also noticed that the best features for judging simplification are different for short sentences and for longer documents.

Overall, our research shows that there is no single, agreed-upon way to measure if a text has been successfully simplified. This means that more work is needed to find better ways to define and measure simplification quality.

10. Bibliographical References

Sweta Agrawal and Marine Carpuat. 2024. Do text simplification systems preserve meaning? a human evaluation via reading comprehension. *Transactions of the Association for Computational Linguistics*, 12:432–448.

Sandra Aluisio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. 2010. Readability assessment for text simplification. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9, Los Angeles, California. Association for Computational Linguistics.

Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. 2020. ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4668–4679, Online. Association for Computational Linguistics.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. The (un)suitability of automatic evaluation metrics for text simplification. *Computational Linguistics*, 47(4):861–889.

Mohammad Alzaid, Faisal R Ali, and Emma Stapleton. 2024. Limitations of readability assessment tools. *European Archives of Oto-Rhino-Laryngology*, 281(9):5021–5022.

Yuki Arase, Satoru Uchida, and Tomoyuki Kajiwara. 2022. CEFR-based sentence difficulty annotation and assessment. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6206–6219, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Abdullah Barayan, Jose Camacho-Collados, and Fernando Alva-Manchego. 2025. Analysing zero-shot readability-controlled sentence simplification. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6762–6781, Abu Dhabi, UAE. Association for Computational Linguistics.

Alessia Battisti, Dominik Pfütze, Andreas Säuberli, Marek Kostrzewa, and Sarah Ebling. 2020. A corpus for automatic readability assessment and

text simplification of German. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3302–3311, Marseille, France. European Language Resources Association.

Rémi Cardon and Adrien Bibal. 2023. On operations in automatic text simplification. In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 116–130, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

John Carroll, Guido Minnen, Darren Pearce, Yvonne Canning, Siobhan Devlin, and John Tait. 1999. Simplifying text for language-impaired readers. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 269–270, Bergen, Norway. Association for Computational Linguistics.

R. Chandrasekar, Christine Doran, and B. Srinivas. 1996. Motivations and methods for text simplification. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

BNC Consortium. 2007. British national corpus 1994. Literary and Linguistic Data Service.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2024. Evaluating document simplification: On the importance of separately assessing simplicity and meaning preservation. *LREC-COLING 2024*, page 1.

Scott A Crossley, Kristopher Kyle, and Mihai Dascalu. 2019. The tool for the automatic analysis of cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior research methods*, 51:14–27.

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

François Daoust, Léo Laroche, and Lise Ouellet. 1996. Sato-calibrage: Présentation d'un outil d'assistance au choix et à la rédaction de textes pour l'enseignement. *Revue québécoise de linguistique*, 25(1):205–234.

Mark Davies. 2008. The corpus of contemporary american english (coca). Available online at https://www.english-corpora.org/coca/.

Maria De Martino. 2023. Processing effort during reading texts in young adults: Text simplification, readability assessment and preliminary eye-tracking data. In *Proceedings of the 9th Italian Conference on Computational Linguistics (CLiC-it 2023)*, pages 179–184, Venice, Italy. CEUR Workshop Proceedings.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. READ–IT: Assessing readability of Italian texts with a view to text simplification. In *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, pages 73–83, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Tovly Deutsch, Masoud Jasbi, and Stuart Shieber. 2020. Linguistic features for readability assessment. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–17, Seattle, WA, USA → Online. Association for Computational Linguistics.

Björn Engelmann, Christin Katharina Kreutz, Fabian Haak, and Philipp Schaer. 2024. ARTS: Assessing readability & text simplicity. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14925–14942, Miami, Florida, USA. Association for Computational Linguistics.

Dengzhao Fang, Jipeng Qiang, Xiaoye Ouyang, Yi Zhu, Yunhao Yuan, and Yun Li. 2025. Collaborative document simplification using multi-agent systems. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 897–912, Abu Dhabi, UAE. Association for Computational Linguistics.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Lorenzo Jaime Flores, Heyuan Huang, Kejian Shi, Sophie Chheang, and Arman Cohan. 2023. Medical text simplification: Optimizing for readability with unlikelihood training and reranked beam search decoding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4859–4873, Singapore. Association for Computational Linguistics.

Peter W Foltz, Walter Kintsch, and Thomas K Landauer. 1998. The measurement of textual coherence with latent semantic analysis. *Discourse processes*, 25(2-3):285–307.

Thomas François. 2015. When readability meets computational linguistics: a new paradigm in readability. *Revue française de linguistique appliquée*, XX(2):79–97.

Gintarė Grigonyte, Maria Kvist, Sumithra Velupillai, and Mats Wirén. 2014. Improving readability of Swedish electronic health records through lexical simplification: First results. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 74–83, Gothenburg, Sweden. Association for Computational Linguistics.

Joey Z Gu, Grayson L Baird, Antonio Escamilla Guevara, Young-Jin Sohn, Melis Lydston, Christopher Doyle, Sarah EA Tevis, and Randy C Miles. 2024. A systematic review and meta-analysis of english language online patient education materials in breast cancer: Is readability the only story? *The Breast*, page 103722.

Robert Gunning. 1952. The technique of clear writing. *McGraw-Hill*.

Reem Hazim, Hind Saddiki, Bashar Alhafni, Muhamed Al Khalil, and Nizar Habash. 2022. Arabic word-level readability visualization for assisted text simplification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 242–249, Abu Dhabi, UAE. Association for Computational Linguistics.

Nam Huong Dau, Duy Van Nguyen, and Hai Thi Thanh Diem. 2024. Annual report readability and firms' investment decisions. *Cogent Economics & Finance*, 12(1):2296230.

Joseph Marvin Imperial and Harish Tayyar Madabushi. 2023. Flesch or fumble? evaluating readability standard alignment of instruction-tuned language models. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 205–223, Singapore. Association for Computational Linguistics.

Zhang Jingshen, Chen Xinglu, Qiu Xinying, Wang Zhimin, and Feng Wenhe. 2024. Readability-guided idiom-aware sentence simplification (RISS) for Chinese. In *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 1: Main Conference)*, pages 1183–1200, Taiyuan, China. Chinese Information Processing Society of China.

Tannon Kew, Alison Chi, Laura Vásquez-Rodríguez, Sweta Agrawal, Dennis Aumiller, Fernando Alva-Manchego, and Matthew Shardlow. 2023. BLESS: Benchmarking large language models on sentence simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13291–13309, Singapore. Association for Computational Linguistics.

J Peter Kincaid, RP Fishburne, RL Rogers, and BS Chissom. 1975. Derivation of new readability formulas (automated reliability index, fog count and flesch reading ease formula) for navy enlisted personnel (research branch report 8-75). memphis, tn: Naval air station; 1975. *Naval Technical Training, US Naval Air Station: Millington, TN*.

Kristopher Kyle, Scott Crossley, and Cynthia Berger. 2018. The tool for the automatic analysis of lexical sophistication (taales): Version 2.0. *Behavior research methods*, 50:1030–1046.

Kristopher Kyle, Scott A Crossley, and Scott Jarvis. 2021. Assessing the validity of lexical diversity indices using direct judgements. *Language Assessment Quarterly*, 18(2):154–170.

Bruce W. Lee, Yoo Sung Jang, and Jason Lee. 2021. Pushing on text readability assessment: A transformer meets handcrafted linguistic features. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10669–10686, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bertha A Lively and Sidney L Pressey. 1923. A method for measuring the vocabulary burden of textbooks. *Educational administration and supervision*, 9(7):389–398.

Xiaofei Lu. 2010. Automatic analysis of syntactic complexity in second language writing. *International journal of corpus linguistics*, 15(4):474–496.

Mounica Maddela and Fernando Alva-Manchego. 2025. Adapting sentence-level automatic metrics for document-level simplification evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6444–6459, Albuquerque, New Mexico. Association for Computational Linguistics.

Mounica Maddela and Wei Xu. 2018. A word-complexity lexicon and a neural readability ranking model for lexical simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3749–3760, Brussels, Belgium. Association for Computational Linguistics.

G Harry Mc Laughlin. 1969. Smog grading-a new readability formula. *Journal of reading*, 12(8):639–646.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Kaijie Mo and Renfen Hu. 2024. ExpertEase: A multi-agent framework for grade-specific document simplification with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9080–9099, Miami, Florida, USA. Association for Computational Linguistics.

Kai North, Tharindu Ranasinghe, Matthew Shardlow, and Marcos Zampieri. 2025. Deep learning approaches to lexical simplification: A survey. *Journal of Intelligent Information Systems*, 63(1):111–134.

John O'hayre. 1966. *Gobbledygook has gotta go*. US Department of the Interior, Bureau of Land Management.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Antonio Flavio Paula and Celso Camilo-Junior. 2024. Evaluating the simplification of Brazilian legal rulings in LLMs using readability scores as a target. In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, pages 117–125, Miami, Florida, USA. Association for Computational Linguistics.

Horacio Saggion. 2017. *Automatic Text Simplification*. Morgan & Claypool Publishers.

Karen Scholz and Markus Wenzel. 2025. Evaluating readability metrics for German medical text simplification. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6049–6062, Abu Dhabi, UAE. Association for Computational Linguistics.

Luo Si and Jamie Callan. 2001. A statistical model for scientific readability. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 574–576.

E.A. Smith and R.J. Senter. 1967. *Automated Readability Index*. AMRL-TR. Aerospace Medical Research Laboratories, Aerospace Medical Division, Air Force Systems Command.

Sanja Štajner and Horacio Saggion. 2013. Readability indices for automatic evaluation of text simplification systems: A feasibility study for Spanish. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 374–382, Nagoya, Japan. Asian Federation of Natural Language Processing.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018a. BLEU is not suitable for the evaluation of text simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 738–744, Brussels, Belgium. Association for Computational Linguistics.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018b. Semantic structural evaluation for text simplification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 685–696, New Orleans, Louisiana. Association for Computational Linguistics.

Renliang Sun, Hanqi Jin, and Xiaojun Wan. 2021. Document-level text simplification: Dataset, criteria and baseline. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7997–8013, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Teerapaun Tanprasert and David Kauchak. 2021. Flesch-kincaid is not a text simplification evaluation metric. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 1–14, Online. Association for Computational Linguistics.

Edward L Thorndike. 1921. Word knowledge in the elementary school. *Teachers College Record*, 22(4):1–27.

Sowmya Vajjala. 2022. Trends, limitations and open challenges in automatic readability assessment research. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5366–5377, Marseille, France. European Language Resources Association.

Sowmya Vajjala and Ivana Lučić. 2018. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 297–304, New Orleans, Louisiana. Association for Computational Linguistics.

Laura Vásquez-Rodríguez, Matthew Shardlow, Piotr Przybyła, and Sophia Ananiadou. 2023. Document-level text simplification with coherence evaluation. In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 85–101, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez,

Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Rodrigo Wilkens, David Alfter, Xiaoou Wang, Alice Pintard, Anaïs Tack, Kevin P. Yancey, and Thomas François. 2022. FABRA: French aggregator-based readability assessment toolkit. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1217–1233, Marseille, France. European Language Resources Association.

Rodrigo Wilkens, Patrick Watrin, Rémi Cardon, Alice Pintard, Isabelle Gribomont, and Thomas François. 2024. Exploring hybrid approaches to readability: experiments on the complementarity between linguistic features and transformers. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2316–2331, St. Julian's, Malta. Association for Computational Linguistics.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Victoria Yaneva, Irina Temnikova, and Ruslan Mitkov. 2016. Evaluating the readability of text simplification output for readers with cognitive disabilities. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 293–299, Portorož, Slovenia. European Language Resources Association (ELRA).

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

A. Appendices

A.1. Correlation Tables

The following tables report the top 10 readability features or metrics by absolute value of significant Pearson correlation coefficient, for each combination of dataset, computation mode (simp = simple side only; delta = difference simple−original) and evaluation criterion.

(a) simp – simplicity

Variable	r
log_ttr_aw	0.199
log_ttr_cw	0.193
McD_CD_FW	0.178
basic_connectives	-0.173
lemma_ttr	0.173
lemma_mattr	0.173
MRC_Familiarity_CW	0.168
msttr50_aw	0.164
mattr50_aw	0.164
bigram_lemma_ttr	0.159

(b) delta – simplicity

Variable	r
conjunctions	0.215
basic_connectives	0.208
av_pobj_deps_NN	-0.178
log_ttr_cw	-0.176
adv_ttr	-0.169
log_ttr_aw	-0.167
av_pobj_deps	-0.159
MRC_Familiarity_CW	-0.153
MRC_Imageability_CW	-0.152
hyper_verb_noun_Sav_P1	-0.150

(c) simp – fluency

Variable	r
COCA_magazine_bi_MI	0.262
log_ttr_aw	0.242
COCA_news_bi_MI	0.237
COCA_fiction_bi_MI	0.237
basic_connectives	-0.218
COCA_spoken_bi_MI	0.217
log_ttr_cw	0.217
conjunctions	-0.203
conj_per_cl	-0.195
acad_lemma_attested	0.195

(d) delta – fluency

Variable	r
root_ttr_aw	-0.300
root_ttr_cw	-0.283
basic_ntypes	-0.242
conjunctions	0.239
mtld_ma_wrap_aw	-0.237
basic_ncontent_types	-0.235
basic_connectives	0.226
av_pobj_deps_NN	-0.212
log_ttr_aw	-0.212
root_ttr_fw	-0.211

(e) simp – meaning

Variable	r
root_ttr_cw	0.295
root_ttr_aw	0.286
log_ttr_cw	0.269
basic_ncontent_types	0.254
hyper_verb_noun_Sav_P1	0.239
mtld_ma_wrap_aw	0.234
hyper_verb_noun_Sav_Pav	0.233
hyper_verb_noun_s1_p1	0.232
linsear	0.231
basic_ntypes	0.231

(f) delta – meaning

Variable	r
root_ttr_aw	-0.426
root_ttr_cw	-0.398
basic_ntypes	-0.392
nwords	-0.390
Word Count	-0.383
basic_ntokens	-0.373
basic_ncontent_tokens	-0.365
mtld_ma_wrap_aw	-0.363
basic_ncontent_types	-0.361
basic_nfunction_types	-0.350

Table 3: Top absolute values of significant correlation coefficients ($p < .001$) between human judgment on the SimplicityDA dataset and readability measures.

(a) simp – simplicity

Variable	r
Word Count	-0.205
Kuperman_AoA_FW	-0.138
Phono_N_FW	0.116
Phono_N_H_FW	0.113
hyper_noun_S1_P1	0.110
hyper_noun_Sav_Pav	0.102
MRC_Familiarity_FW	0.096
COCA_fiction_Range_CW	-0.095
BNC_Spoken_3gram_NF	-0.093
poly_adj	-0.091

(b) delta – simplicity

Variable	r
Word Count	0.298
TL_Freq_FW_Log	-0.267
COCA_fiction_Frequency_Log_FW	-0.258
KF_Freq_FW_Log	-0.252
BNC_Written_Freq_FW_Log	-0.249
COCA_news_Frequency_Log_FW	-0.245
COCA_magazine_Frequency_Log_FW	-0.239
AWL_Sublist_5_Normed	0.237
BNC_Spoken_Freq_FW_Log	-0.236
Brown_Freq_FW_Log	-0.234

(c) simp – fluency

Variable	r
COCA_spoken_Trigram_Frequency_Log	0.103
WN_SD_CW	-0.101
COCA_news_tri_2_DP	0.099
Brysbaert_CC_AW	0.098
COCA_magazine_tri_2_MI2	0.098
COCA_magazine_tri_2_DP	0.098
COCA_spoken_tri_prop_20k	0.098
OG_N_H	-0.098
Freq_N_OGH	-0.098
COCA_news_tri_prop_10k	0.096

(d) delta – fluency

Variable	r
AWL_Sublist_10_Normed	-0.291
COCA_fiction_Frequency_FW	-0.216
BNC_Spoken_3gram_NF_Log	0.214
TL_Freq_FW	-0.202

(e) simp – meaning

Variable	r
Word Count	-0.433
Kuperman_AoA_FW	-0.243
Brown_Freq_CW	0.240
TL_Freq_CW	0.239
KF_Freq_CW	0.219
OLDF_FW	0.213
Freq_N_OG_CW	0.211
OG_N_H_CW	0.200
Freq_N_OGH_CW	0.200
poly_adj	-0.192

(f) delta – meaning

Variable	r
Word Count	0.342
Kuperman_AoA_AW	0.270
COCA_spoken_Frequency_Log_CW	-0.255
PLDF_FW	-0.248
COCA_spoken_RL_CW	-0.228
SUBTLEXus_Range_FW	-0.225
COCA_news_RL_FW	-0.224
COCA_spoken_RL_FW	-0.221
KF_Ncats_FW	-0.220
Kuperman_AoA_CW	0.211

Table 4: Top absolute values of significant correlation coefficients ($p < .05$) between human judgment on the DWiki dataset and features.

(a) simp – bertscore_F1

Variable	r
root_ttr_aw	0.339
rootTTRCW	0.317
log_ttr_cw	0.311
mtld_ma_wrap_aw	0.294
hyper_verb_noun_Sav_P1	0.291
hyper_verb_noun_Sav_Pav	0.283
hyper_verb_noun_s1_p1	0.279
log_ttr_aw	0.277
hyper_noun_S1_P1	0.276
basic_ntypes	0.273

(b) delta – bertscore_F1

Variable	r
root_ttr_aw	-0.518
basic_ntypes	-0.466
root_ttr_cw	-0.464
nwords	-0.461
mtld_ma_wrap_aw	-0.457
Word Count	-0.445
basic_ncontent_tokens	-0.434
basic_ncontent_types	-0.432
basic_ntokens	-0.429
linsear	-0.413

(c) simp – bleu

Variable	r
root_ttr_aw	0.352
root_ttr_cw	0.341
linsear	0.329
mtld_ma_wrap_aw	0.324
basic_ntypes	0.315
Word Count	0.314
nwords	0.313
basic_ncontent_types	0.303
log_ttr_cw	0.295
fkgl	0.290

(d) delta – bleu

Variable	r
hyper_verb_noun_Sav_P1	-0.403
hyper_verb_noun_Sav_Pav	-0.384
hyper_verb_noun_s1_p1	-0.362
hyper_noun_Sav_P1	-0.305
av_pobj_deps_NN	-0.302
log_ttr_cw	-0.293
av_pobj_deps	-0.291
hyper_noun_S1_P1	-0.291
KF_Freq_CW	0.284
COCA_fiction_Freq_CW	0.282

(e) simp – samsa

Variable	r
cl_ndeps_std_dev	-0.309
basic_ntokens	-0.302
Word Count	-0.285
basic_ntypes	-0.284
basic_nfunction_tokens	-0.277
nwords	-0.277
basic_ncontent_tokens	-0.271
mtld_ma_wrap_aw	-0.265
basic_nfunction_types	-0.261
basic_ncontent_types	-0.253

(f) delta – samsa

Variable	r
COCA_news_RL_AW	0.199
fog	-0.198
COCA_news_RL_CW	0.197
basic_ncontent_types	-0.194
arindex	-0.193
basic_ncontent_tokens	-0.192
COCA_spoken_RL_AW	0.188
mtld_ma_wrap_aw	-0.185
fkgl	-0.184
poly_verb	0.184

(g) simp – sari

Variable	r
root_ttr_aw	0.378
Word Count	0.378
nwords	0.376
root_ttr_cw	0.365
basic_ntypes	0.363
mtld_ma_wrap_aw	0.357
basic_ntokens	0.355
basic_ncontent_types	0.337
basic_ncontent_tokens	0.327
mtld_ma_wrap_cw	0.322

(h) delta – sari

Variable	r
nwords	-0.499
Word Count	-0.496
root_ttr_aw	-0.477
basic_ntypes	-0.461
basic_ntokens	-0.459
mtld_ma_wrap_aw	-0.449
basic_nfunction_types	-0.447
basic_nfunction_tokens	-0.442
root_ttr_cw	-0.420
basic_ncontent_tokens	-0.410

Table 5: Top absolute values of significant correlation coefficients ($p < .05$) between human judgments and automatic metrics, on the SimplicityDA dataset.

(a) simp – D-SARI

Variable	r
Word Count	-0.220
MRC_Imageability_FW	0.172
Brysbaert_CC_FW	0.171
MRC_Concreteness_FW	0.164
MRC_Meaningfulness_FW	0.148
COCA_academic_tri_2_DP	-0.146
KF_Freq_CW	0.146
Kuperman_AoA_FW	-0.134
Brysbaert_CC_AW	0.133
eat_tokens	-0.129

(b) delta – D-SARI

Variable	r
WN_Zscore_CW	0.305
COCA_spoken_tri_MI2	-0.303
WN_Zscore	0.298
COCA_spoken_tri_MI	-0.283
COCA_spoken_Trigram_Range_Log	0.266
WN_Mean_RT_CW	0.263
COCA_spoken_tri_2_MI2	-0.258
Ortho_N_CW	-0.254
WN_Mean_RT	0.252
PLD	0.251

(c) simp – LENS

Variable	r
Word Count	-0.347
McD_CD_FW	0.199
Kuperman_AoA_FW	-0.196
lsa_average_all_cosine	0.191
Brown_Freq_CW	0.188
COCA_magazine_Range_Log_AW	-0.186
Brysbaert_CC_FW	0.186
COCA_academic_tri_2_DP	-0.185
OG_N_H_FW	0.185
Freq_N_OGH_FW	0.185

(d) delta – LENS

Variable	r
COCA_spoken_tri_2_MI	-0.279
COCA_spoken_tri_2_MI2	-0.267
LD_Mean_RT_SD	0.247
LD_Mean_RT_SD_CW	0.241
COCA_fiction_Frequency_AW	0.241
COCA_news_Frequency_AW	0.235
COCA_spoken_tri_MI2	-0.234
COCA_magazine_Frequency_AW	0.226
Brown_Freq_CW_Log	-0.221
poly_noun	0.217

(e) simp – BERTScore_F1

Variable	r
WN_Mean_Accuracy_CW	-0.143
WN_Mean_Accuracy	-0.134
COCA_Fiction_Trigram_Range_Log	0.133
LD_Mean_Accuracy_CW	-0.131
COCA_fiction_tri_2_MI	-0.127
COCA_fiction_tri_2_MI2	-0.112
LD_Mean_Accuracy	-0.112
COCA_spoken_Trigram_Range_Log	0.096
LD_Mean_RT_Zscore	0.092
BNC_Written_Trigram_Freq_Normed_Log	0.090

(f) delta – BERTScore_F1

Variable	r
AWL_Sublist_10_Normed	0.219
COCA_academic_tri_T	-0.202
COCA_magazine_tri_2_T	-0.194
COCA_academic_tri_2_T	-0.192
COCA_news_tri_T	-0.188
COCA_magazine_tri_T	-0.186
COCA_news_tri_2_DP	0.186
COCA_news_tri_2_T	-0.184
COCA_Academic_Trigram_Frequency_Log	-0.181
LD_Mean_RT_SD_CW	-0.179

Table 6: Top absolute values of significant correlation coefficients ($p < .05$) between human judgments and automatic metrics, on the D-WIKI dataset.

Language Proficiency as a Recoverable Dimension in Multilingual LLM Embeddings

Rodrigo Wilkens
University of Exeter
r.wilkens@exeter.ac.uk

Abstract

Understanding whether proficiency is encoded as structured knowledge rather than inferred from surface correlates is critical for interpreting and applying LLMs in educational contexts. We investigate whether multilingual large language model (LLM) embeddings encode language proficiency as a structured recoverable dimension rather than merely supporting predictive classification. Using the UniversalCEFR benchmark, which spans 13 languages and the full proficiency range from A1 to C2, we evaluate the frozen LLM embedding space in two complementary ways. First, we test whether proficiency levels can be predicted directly from frozen embeddings across languages and model variants. The results show that embeddings without task-specific fine-tuning consistently support CEFR classification. Variation in results is strongly associated with the amount of annotated data and language family, suggesting that data availability and cross-linguistic structure matter more than architectural differences. Second, we examine how CEFR levels are organized inside embedding space. We find that texts from lower to higher proficiency levels align along a consistent ordered direction, with higher levels systematically positioned further along this gradient. Distances between levels increase proportionally to their ordinal gap (e.g., A1 vs. C2 is farther apart than B1 vs. B2), indicating a "consistent continuous gradient with overlapping adjacent levels rather than arbitrary clusters. Together, these findings show that CEFR is not only predictable from multilingual LLM embeddings but is also internally structured as an ordered representational dimension.

Keywords: Language Proficiency Assessment, Representation Probing, Multilingual LLM Embeddings, CEFR Modeling

1. Introduction

Language proficiency assessment plays a central role in education, large-scale testing, and increasingly in natural language processing (NLP) applications that support personalized learning and automated feedback (Shermis and Burstein, 2003; Yannakoudakis et al., 2011; Ke and Ng, 2019). Within this context, the Common European Framework of Reference for Languages (CEFR) has become a widely adopted standard for describing learner proficiency across languages and educational systems (of Europe, 2020; of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division, 2001). Automated essay scoring (AES) and automatic readability assessment (ARA) constitute two major strands of computational research concerned with modeling language proficiency. AES focuses on the assessment of learner-produced texts, aiming to evaluate writing quality and proficiency level (Page, 1966; Attali and Burstein, 2006; Yannakoudakis et al., 2011; Ke and Ng, 2019). In contrast, ARA addresses the estimation of text difficulty and its suitability for readers at different proficiency stages (Collins-Thompson, 2014; Vajjala, 2022). Despite their distinct theoretical orientations (i.e., production in the case of AES and reception in the case of ARA), both lines of work have historically operationalized proficiency through observable properties of text. Early approaches relied on linguistic features such as lexical diversity, syntactic complexity, and cohesion measures (Zesch et al., 2015; Ke and Ng, 2019).

Deep learning marked a shift from explicit feature engineering to representation learning. Neural models based on CNNs and LSTMs demonstrated improved ability to capture semantic and discourse-level patterns (Taghipour and Ng, 2016; Alikaniotis et al., 2016; Dong and Zhang, 2016), but struggled with long-range dependencies and scalability. Transformer-based encoders, such as BERT (Devlin et al., 2019), have since become dominant in NLP, enabling fine-tuning for proficiency level prediction tasks (Mayfield and Black, 2020; Rodriguez et al., 2019). Hybrid models that combine contextual embeddings with handcrafted linguistic features have achieved further gains in both AES and ARA (Li et al., 2022; Faseeh et al., 2024; Liu and Lee, 2023; Wilkens et al., 2024).

Generative decoder-based large language models (LLMs) have reshaped the landscape of NLP. Instead of fine-tuning encoder architectures, researchers increasingly rely on prompt engineering, in-context learning, and instruction tuning of autoregressive models (Brown et al., 2020; Chung et al., 2024). This progression reflects a steady expansion of representational capacity and methodological flexibility in automated assessment. In particular, it is still unknown whether the CEFR levels are internally encoded in the models (i.e., corre-

Proceedings of the Joint Workshop on Readability and Text Simplification (READIxTSAR) @ LREC 2026, pages 227–242
11 May 2026.

spond to a coherent latent geometric direction in LLM embedding spaces) or whether the observed predictive success stems from superficial correlations (e.g., text length and lexical frequency). While probing studies have shown that linguistic properties can be recovered from contextual embeddings (Tenney et al., 2019; Hewitt and Manning, 2019; Pimentel et al., 2020), it remains unclear how higher-level constructs such as language proficiency are structured within multilingual embedding spaces. From an educational perspective, understanding whether proficiency is internally structured has direct implications. If proficiency corresponds to a recoverable and continuous representational dimension, it enables new forms of model usage beyond classification, including controllable text adaptation, calibrated feedback generation, and alignment between model representations and pedagogical scales. Conversely, if predictions rely primarily on superficial correlates, their interpretability and pedagogical validity remain limited.

This study addresses that gap by investigating whether frozen multilingual LLM embeddings encode recoverable CEFR structure. To achieve this goal, we formulate two research questions:

1. To what extent do frozen multilingual LLM embeddings support CEFR classification across languages and modeling configurations?

2. Does CEFR proficiency correspond to a recoverable latent geometric axis in embedding space that generalizes across languages and task modalities?

Using multilingual CEFR corpora, we conduct predictive evaluation alongside cross-validated geometric analysis of embedding space to investigate whether CEFR proficiency corresponds to a recoverable latent structure. Our findings indicate that while embeddings provide a multilingual proficiency signal, CEFR levels can be interpreted as continuous latent gradients embedded within higher-dimensional linguistic structure. Our findings contribute to the current discussions on whether LLMs capture linguistically valid proficiency constructs or merely exploit surface correlates.

The remainder of the paper reviews related work on automated assessment and representational analysis (Section 2), introduces the predictive and geometric methodology (Section 3), presents our findings addressing multilingual classification and latent proficiency structure (Section 4), and concludes with an analysis of representational implications and future directions (Sections 5-6).

2. Related Work

2.1. Language Proficiency Assessment

Automated Essay Scoring (AES), concerned with modeling learner language production, and automatic readability assessment (ARA), focused on estimating text difficulty for language reception, have historically relied on handcrafted linguistic indicators such as lexical diversity, syntactic complexity, discourse features, and surface statistics (Page, 1966; Attali and Burstein, 2006; Persing et al., 2010; Zesch et al., 2015; Collins-Thompson, 2014; François and Fairon, 2012). In both traditions, proficiency or difficulty has been operationalized through observable linguistic properties derived from text. Cross-lingual readability studies have further demonstrated that complexity modeling varies across languages and typological settings (Vajjala and Rama, 2018), highlighting the interaction between linguistic structure and proficiency prediction. Neural architectures based on CNNs and LSTMs shifted the focus toward representation learning (Alikaniotis et al., 2016; Taghipour and Ng, 2016; Dong and Zhang, 2016), reducing explicit feature engineering while retaining predictive evaluation as the primary objective. Transformer encoders (Devlin et al., 2019) subsequently became dominant, and fine-tuning approaches were widely adopted for AES and CEFR prediction (Rodriguez et al., 2019; Mayfield and Black, 2020). Hybrid architectures combining contextual embeddings with handcrafted linguistic and readability features consistently achieved strong performance, reinforcing the complementarity between structured linguistic metrics and distributed representations (Li et al., 2022; Faseeh et al., 2024; Liu and Lee, 2023; Wilkens et al., 2024). Across these methodological developments, research in both AES and ARA has remained primarily evaluation-driven, emphasizing predictive metrics over the analysis of representational structure.

2.2. Encoder and Decoder Models in Assessment

Transformer-based encoders dominated early LLM-based assessment through supervised fine-tuning (Devlin et al., 2019; Liu et al., 2019), while multilingual variants demonstrated strong cross-lingual transfer capabilities (Conneau et al., 2020). However, empirical evidence shows that multilingual performance varies with data size, domain, and language similarities (Lauscher et al., 2020; Bjerva et al., 2019).

Advances in autoregressive decoder models have shifted attention toward prompt-based and instruction-tuned evaluation (Brown et al., 2020; Chung et al., 2024). Decoder-based scoring can

achieve competitive results without task-specific fine-tuning, challenging the dominance of encoder-centric pipelines (Seßler et al., 2025; Yancey et al., 2023; Mizumoto and Eguchi, 2023; Liew and Tan, 2024). Nevertheless, comparative studies of encoder and decoder configurations largely assess the system's predictive accuracy.

Probing research demonstrates that linguistic properties can be linearly recovered from contextual embeddings (Tenney et al., 2019; Hewitt and Manning, 2019; Pimentel et al., 2020), and geometric analyses reveal that syntactic and semantic structure often corresponds to identifiable directions or subspaces in high-dimensional representation space (Mu and Andreas, 2020). While such studies show that structured linguistic information is embedded in transformer representations, it primarily targets discrete grammatical or lexical properties. Complex, socially defined constructs such as language proficiency, particularly under standardized frameworks like CEFR, have not been examined as latent geometric gradients in multilingual embedding spaces.

3. Methodology

Aiming to investigate whether multilingual LLM embeddings encode CEFR proficiency as recoverable information, we combine predictive evaluation (RQ1) with geometric analysis (RQ2). The former evaluates classification performance across modeling configurations, while the latter examines whether proficiency corresponds to a latent geometric axis in embedding space.

3.1. Predictive Framework

We evaluate whether frozen multilingual LLM embeddings encode sufficient signal to support CEFR classification across languages. This step establishes representational adequacy before examining geometric structure. We focus on EuroLLM (Martins et al., 2025b,a), a multilingual large language model trained with explicit European-language coverage, as it provides controlled proportions of pretraining across languages. It also supports both base (enriched) and instruction-tuned (instruct-enriched) variants.

Using EuroLLM allows us to directly examine whether differences in instruction tuning or pretraining exposure influence CEFR encoding across languages.[1] The comparison isolates the effect of instruction tuning. Otherwise, similar performance would suggest that the proficiency signal emerges during the base model's training. The information encoded in the model has been associated with scaling effects (Brown et al., 2020). Therefore, we evaluate two model sizes: 1.7B and 9B parameters. All models are run in frozen mode, no fine-tuning is performed, and the model input is exclusively the text document without additional instructions. This isolates pretrained representational capacity from task-specific adaptation. We extract embeddings using the HuggingFace library (Wolf et al., 2019) and selecting the representations from the final hidden layer.

To evaluate the separability of the proficiency signal, we train three downstream classifiers to predict the 6-CEFR levels (A1-C2): Logistic Regression (LR), Linear Support Vector Machine (SVM) with a linear kernel, and Multi-layer Perceptron (MLP). LR and SVM are used to test whether the CEFR signal is linearly separable in the embedding space.[2] The MLP introduces controlled non-linearity; thus, a performance improvement suggests that the CEFR signal is present but not linearly organized. These three models were trained using stratified N-fold cross-validation[3]. Models were evaluated using accuracy, macro F1, Quadratic Weighted Kappa (QWK), and Mean Absolute Error (MAE).[4]

Moving beyond machine learning model ranking, we investigate what drives variation in CEFR classification performance across languages. Therefore, we perform a secondary analysis at the level of language-corpus configurations. Concretely, after computing cross-validated performance scores (e.g., Macro F1) for each combination of language, corpus, model variant (Enriched vs Instruct-Enriched), model size (1.7B vs 9B), and downstream classifier, we treat each such configuration as a single data point in a regression analysis. In other words, for every language-corpus pair and modeling setup, we obtain one performance score. These scores constitute the dependent variable in our analysis. We estimate a series of linear regression models using Ordinary Least Squares (OLS). In this context, OLS estimates the average effect of each explanatory variable on classification performance while holding other variables constant. The resulting coefficients can be interpreted as the expected change in performance associated with a

[1] We focus on a single model family to maintain a controlled multilingual training setting, allowing us to isolate representational effects without confounding differences in architecture or training data.

[2] Linear probes are standard in representation analysis because they provide a minimal-capacity test of recoverability (Hewitt and Manning, 2019; Pimentel et al., 2020).

[3] The N is determined by dataset size: 5-fold CV when fewer than 500 instances, 10-fold CV when between 500 and 5,000 instances, and 5-fold CV in the other cases.

[4] QWK is included due to its ordinal sensitivity and established use in AES research (Taghipour and Ng, 2016). MAE provides complementary information regarding average ordinal deviation.

one-unit change in the predictor. To account for variability across languages, we also use mixed-effects models where language is treated as a random effect. Intuitively, this allows us to separate general trends (e.g., dataset size effects) from language-specific variation, ensuring that observed effects are not driven by particular languages but reflect broader patterns. At each step, we examine the increase in explained variance (R^2) to determine how much additional information each block of variables contributes. This incremental approach allows us to assess whether performance differences are primarily driven by architecture, pretraining exposure, supervision volume, or broader linguistic structure.

3.2. Representational Framework

Given our goal of testing whether proficiency can be recovered as a coherent and ordered geometric factor, we move beyond predictive accuracy and examine the internal structure of embedding space. In this sense, a purely unsupervised analysis would identify directions of maximal variance, which in multilingual embeddings are typically driven by language identity, topic distribution, or corpus effects. Proficiency, however, is not expected to dominate global variance. Therefore, the absence of visible clustering in an unsupervised projection would not imply absence of structured proficiency encoding. For this reason, we adopt a supervised probing strategy. Rather than asking which directions explain the most variance, we directly test whether a simple linear direction can be learned that aligns embeddings with CEFR levels. If proficiency is encoded in representation space, it should be recoverable through such a projection even if it does not correspond to a principal component.

For each LLM variant, we train a Ridge regression model to predict numeric CEFR levels from standardized document embeddings.[5] The learned regression coefficients define a direction in embedding space that best aligns with proficiency. Each document is then projected onto this direction, yielding a scalar value that represents its position along the candidate proficiency axis. To ensure generalizability, the direction is learned using cross-validation. A linear probe is intentionally chosen as a minimal-capacity test (Tenney et al., 2019; Hewitt and Manning, 2019). If CEFR corresponds to structured information encoded in embeddings, it should be recoverable without requiring complex nonlinear transformations.

If CEFR is encoded as an ordered geometric factor, projected values should increase with proficiency level. We quantify this alignment using Spearman rank correlation between projection values and CEFR labels.[6] Spearman's ρ is appropriate given the ordinal nature of CEFR. A high correlation indicates a consistent ordering of levels along the recovered axis.

Alignment alone does not guarantee meaningful geometric separation. A projection may correlate strongly with CEFR while still exhibiting substantial overlap between adjacent levels. We therefore examine dispersion patterns along the recovered axis. For each CEFR level, we compute: (1) the mean projection value (centroid), and (2) the variance of projection values within that level. We then quantify CEFR-level separation using the Fisher ratio, defined as the average squared distance between level centroids divided by the average within-level variance. This ratio evaluates whether between-level differences dominate internal variability. A high Fisher value indicates that CEFR levels form compact and well-separated bands along the axis, rather than diffuse, overlapping distributions.

To further evaluate ordinal geometry, we compute pairwise distances between level centroids and examine their relationship by measuring how far apart two CEFR levels are in ordinal terms. If proficiency is encoded as a continuous gradient, centroid distance should increase as the ordinal difference between levels increases.

To contextualize these findings, we also examine unsupervised principal component projections derived solely from embedding variance, without using CEFR labels or the learned proficiency axis. If proficiency were a dominant global variance component, CEFR levels would align along the leading principal components. By comparing unsupervised structure with supervised projections, we determine whether proficiency is a principal variance driver or a recoverable latent dimension embedded within higher-dimensional structure.

3.3. Corpus

We conduct our experiments on UniversalCEFR (Imperial et al., 2025), a large-scale multilingual benchmark for CEFR-based proficiency modeling. This dataset is composed of CEFR-annotated texts spanning 13 languages and covering the full proficiency spectrum from A1 to C2. Table 1 summarizes the number of documents per language in our experimental setup.

UniversalCEFR integrates both learner-produced texts and pedagogically curated reference materi-

[5] We use ridge regression to estimate a single linear projection from embedding space to ordinal CEFR levels. Unlike multi-class classification models, ridge regression yields a continuous direction that is interpretable as a proficiency axis. L2 regularization further ensures stability given the high dimensionality and collinearity of LLM embeddings.

[6] To assess robustness, we compute 95% confidence intervals using bootstrap resampling over documents.

Language	# Docs	CEFR Levels
Arabic (ar)	2,160	A1–C2
Czech (cz)	441	A1–C1
Welsh (cy)	1,372	A1–A2
German (de)	1,542	A1–C2
English (en)	15,513	A1–C2
Spanish (es)	31,355	A1–C2
French (fr)	2,013	A1–C2
Hindi (hi)	1,491	A1–C1
Italian (it)	813	A1–B2
Dutch (nl)	3,596	A1–C2
Portuguese (pt)	1,423	A1–C2
Russian (ru)	1,758	A1–C2

Table 1: Number of CEFR-annotated documents per language in our experimental setup.

als under a unified CEFR labeling scheme. The languages represented belong to diverse typological families, including Germanic, Romance, Slavic, Indo-Aryan, Celtic, and Semitic languages. This typological and task diversity provides a heterogeneous yet controlled environment for multilingual proficiency analysis.

UniversalCEFR is particularly suitable for our study because its large scale and full CEFR coverage enable robust estimation of proficiency-related structure across all six levels. Its multilingual composition allows us to test whether proficiency constitutes a coherent representational dimension across typologically diverse languages. Furthermore, the inclusion of both learner and reference texts enables the analysis of production- and reception-oriented proficiency within a shared embedding space. Together, these properties make UniversalCEFR an appropriate benchmark for evaluating both predictive recoverability (RQ1) and geometric coherence (RQ2).

4. Results

4.1. Predictive Performance

Table 2 reports cross-validated performance for each model configuration across languages. Frozen embeddings from all EuroLLM variants support CEFR classification at levels exceeding the prompt-based approach reported in Imperial et al. (2025).[7]

Performance differences between Enriched and Instruct-Enriched are small and inconsistent across metrics. Similarly, the difference between 1.7B and 9B models is modest. These results indicate that CEFR-relevant signal is already present in pretrained embeddings and does not depend strongly on model variant.

Across classifiers, Logistic Regression and linear SVM perform comparably to the MLP. The limited improvement from the nonlinear MLP suggests that CEFR signal is largely linearly recoverable from embeddings.

The relationship between pretraining proportion and CEFR performance reveals no meaningful association. The Spearman correlation between pretraining proportion and F1 is negligible ($\rho = -0.012$). Although pretraining proportion appears statistically significant in intermediate regression models, its effect disappears once language family is included. Pretraining exposure alone, therefore, does not robustly predict CEFR performance. Figure 1 provides a visual illustration of this pattern.

In contrast, dataset size exhibits a consistent association with performance. Adding log-transformed dataset size increases explained variance from $R^2 = 0.017$ to $R^2 = 0.139$ ($\Delta R^2 = 0.122$). In OLS models, the coefficient for dataset size remains positive and statistically significant. In mixed-effects models controlling for language-level, the effect is no longer statistically significant.

Language family accounts for a substantial portion of performance variability. In OLS regression, introducing language family increases explained variance to $R^2 = 0.49$ (Adj. $R^2 = 0.455$), representing the largest improvement across model specifications. Romance and Celtic languages exhibit significantly higher performance relative to the reference category. To account for non-independence across languages, we further estimate a mixed-effects model with language as a random factor. In this specification, family-level patterns remain observable, although effect sizes are attenuated. This indicates that part of the variance attributed to language family reflects language-specific structure rather than purely genealogical grouping. Taken together, the regression results indicate that dataset size and language family explain substantially more variance than model configuration or pretraining proportion.[8]

4.2. Representational Structure

Given the capacity of simple machine learning models to predict the CEFR level using embeddings from a frozen model, we now examine whether CEFR is encoded as a coherent latent direction in embedding space.

Table 3 shows that all three models yield strong ordinal alignment between projection values and CEFR levels, with cross-validated Spearman correlations above 0.89. The narrow bootstrap confidence intervals indicate that this alignment is sta-

[7] We consider the best F1 using prompt-based method reported on Imperial et al. (2025) for the EuroLLM9B.

[8] Performance distributions across languages and families are illustrated in Appendix Figure 7.

Lang	Acc	F1	QWK	MAE	σ_{F1}	Prompt F1
es	0.986	0.986	0.987	0.025	0.302	0.28
cy	0.959	0.959	0.914	0.041	0.035	0.26
it	0.843	0.627	0.740	0.157	0.034	0.42
pt	0.613	0.567	0.696	0.601	0.177	0.21
de	0.681	0.563	0.815	0.354	0.125	0.38
cs	0.733	0.535	0.762	0.270	0.018	0.33
ar	0.495	0.461	0.661	0.616	0.068	0.35
en	0.516	0.421	0.696	0.554	0.106	0.23
fr	0.438	0.402	0.643	0.711	0.044	0.28
nl	0.443	0.395	0.553	0.665	0.008	0.32
ru	0.425	0.395	0.742	0.760	0.037	0.21
hi	0.367	0.366	0.659	0.961	0.021	0.21

Table 2: Weighted performance by language. Metrics are averaged across all model configurations, with dataset size as the weight. σ_{F1} reports the standard deviation of F1 across configurations. Table 4 (Appendix) shows the results for all model configurations.

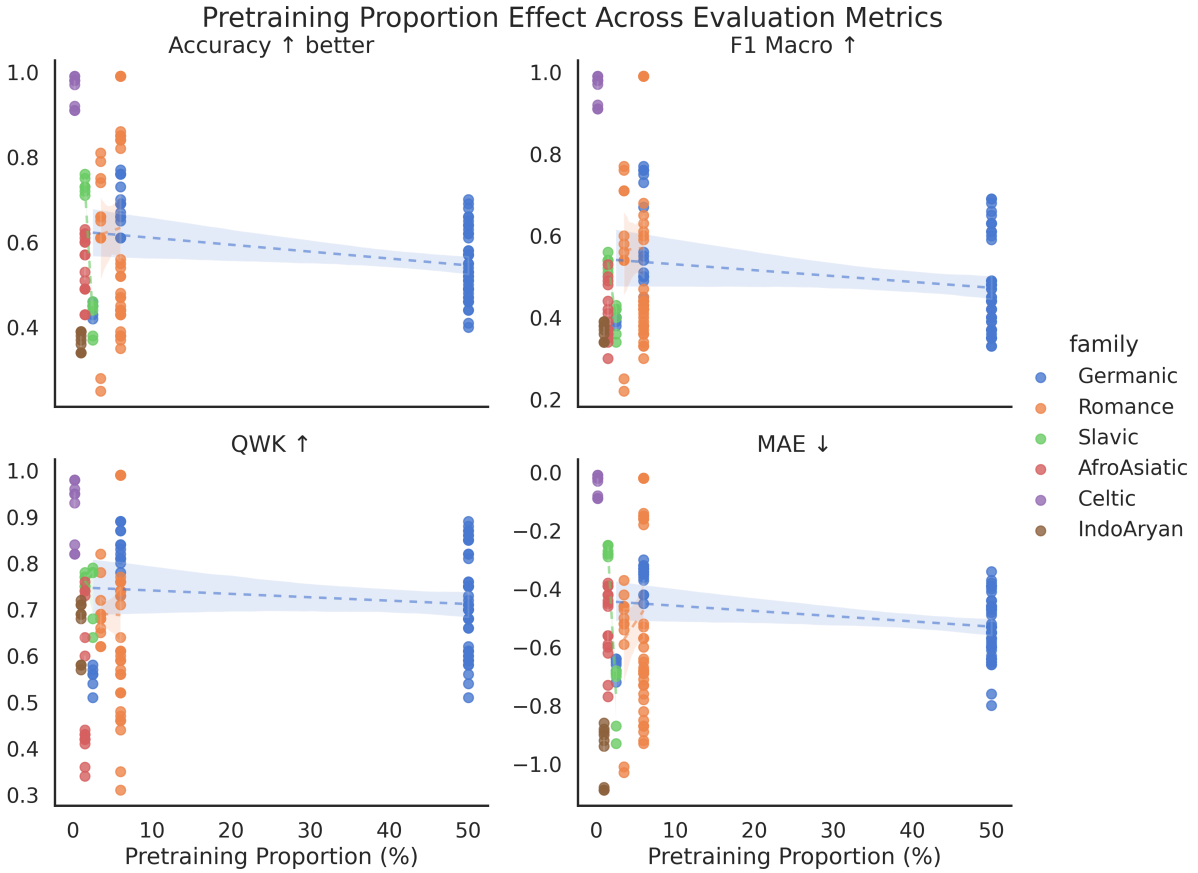


Figure 1: Relationship between pretraining proportion and CEFR classification performance across metrics.

ble and not driven by sampling variability. The 9B model achieves the highest correlation ($\rho_{CV} = 0.902$) and the largest Fisher separation ratio (12.30), suggesting slightly sharper geometric separation of proficiency bands. However, differences across model variants are modest. These results indicate that CEFR is not merely recoverable through classification (RQ1) but is encoded as a coherent geometric direction within embedding space. The high Fisher ratios further suggest structured banding rather than random overlap.

To visually ground these quantitative results, cross-validated projections onto the learned proficiency axis show monotonic increases in projection values from A1 to C2 across models. Although adjacent levels partially overlap, global ordinal struc-

Model	ρ_{CV}	95% CI	Fisher Ratio
1.7B	0.895	[0.893, 0.897]	11.24
1.7B-Instruct	0.896	[0.894, 0.898]	11.26
9B	0.902	[0.900, 0.904]	12.30

Table 3: Cross-validated ordinal alignment and geometric separation metrics.

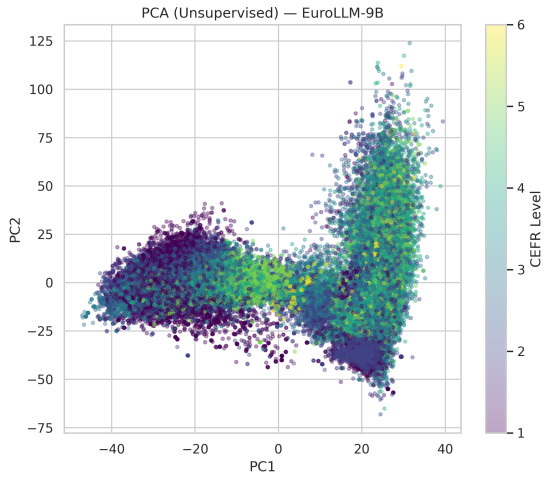


Figure 2: Unsupervised PCA projection (EuroLLM-9B example).

ture is preserved. Examining centroid geometry further reveals that between-level distances increase approximately in proportion to ordinal differences, supporting the interpretation of CEFR as a continuous representational gradient rather than a set of discrete clusters.[9]

Figure 2 shows that CEFR levels do not align with dominant variance directions. The absence of an ordered gradient indicates that proficiency is not a primary organizing factor in embedding space. In contrast, Figure 3 shows the same embeddings reorganized using the supervised proficiency direction. A clear horizontal gradient from A1 to C2 becomes visible, while orthogonal dispersion remains substantial. The contrast between Figures 2 and 3 indicates that proficiency does not dominate global variance but is recoverable as a task-relevant latent dimension.

Finally, Figure 3 shows that both task types align along the same horizontal proficiency axis. At the same time, in both space representations, two vertically separated modality clusters (i.e., learner vs. reference texts) emerge along the orthogonal dimension.[10]. This indicates that the recovered direction captures shared proficiency-related structure,

[9]Cross-validated axis projections and centroid distance diagnostics are provided in Appendix Figures 4 and 5.

[10]Learner and reference clusters are illustrated in Figure 6.

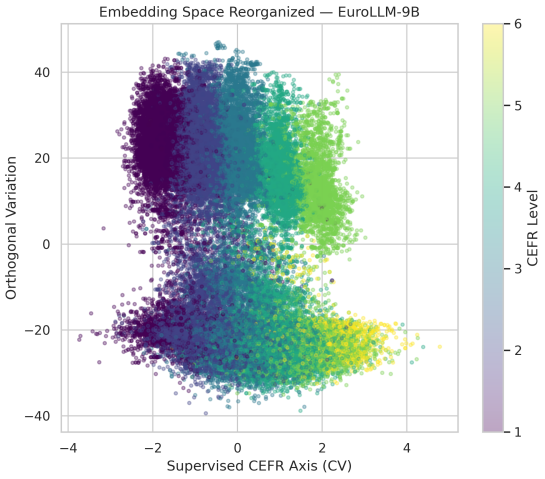


Figure 3: Embedding space reorganized using the supervised CEFR axis.

while modality-specific variation is expressed primarily in dimensions orthogonal to the proficiency axis.

5. Discussion

Our results show that frozen multilingual LLM embeddings consistently support CEFR classification across languages. This finding extends prior work in Automated Essay Scoring (AES) and readability assessment, where encoder-based models typically require task-specific fine-tuning or feature augmentation to achieve competitive performance (Taghipour and Ng, 2016; Alikaniotis et al., 2016).

We demonstrate that pretrained embeddings alone (without fine-tuning or linguistic features) already encode a signal for multilingual CEFR prediction. This challenges the prevailing assumption that assessment tasks necessarily require task-adapted architectures or sophisticated prompts. Instead, our findings suggest that CEFR-relevant linguistic structure emerges during large-scale pretraining.

Importantly, performance differences between Enriched and Instruct-Enriched variants are minimal. This indicates that CEFR signal does not primarily arise from instruction-following alignment, but is already present in base pretrained representations.

A central finding concerns the limited explanatory power of pretraining proportion. While generative LLM literature often emphasizes scaling and data exposure as primary drivers of performance (Brown et al., 2020), our regression analyses show that pretraining proportion does not robustly predict CEFR classification once language family is controlled.

In contrast, supervision volume (dataset size) and linguistic grouping explain substantially more variance. This aligns with multilingual NLP re-

search showing that cross-lingual transfer depends strongly on typological proximity and annotation density (Lauscher et al., 2020; Bjerva et al., 2019).

For CEFR assessment, this implies that increasing annotated training data may yield greater gains than modifying embedding architectures or increasing raw pretraining exposure.

The results provide evidence that CEFR is encoded as a recoverable dimension. High cross-validated ordinal correlations and substantial Fisher separation ratios indicate that proficiency is not merely classifiable.

Proficiency does not align with dominant principal components. This shows that CEFR is not a primary global variance factor (such as language or topic), but a latent dimension embedded within a higher-dimensional structure. This distinction reframes CEFR modeling: rather than asking whether models can predict proficiency, we show that proficiency corresponds to an internally organized representational gradient.

The approximately linear relationship between centroid distance and ordinal gap suggests that CEFR is encoded as a continuous proficiency gradient rather than as sharply separated categories. Adjacent levels overlap substantially, while global ordering remains stable. This geometric perspective aligns with theoretical interpretations of CEFR as a continuum of communicative competence rather than as strictly discrete bands.

The recovered proficiency axis is shared across learner (production) and reference (reception) texts. While learner texts exhibit greater orthogonal dispersion, particularly at lower levels, the horizontal ordering remains consistent.

These findings point to a recovered dimension that captures shared linguistic complexity rather than task-specific artefacts. Production variability manifests primarily in orthogonal directions, while proficiency alignment remains stable. These findings are consistent with an interpretation in terms of linguistic competence rather than purely corpus-specific conventions.

6. Conclusion

This study investigated whether multilingual LLM embeddings encode CEFR proficiency as recoverable and structured information. Investigating the extent to which frozen multilingual LLM embeddings support CEFR classification across languages and modeling configurations (*RQ1*), we showed that frozen EuroLLM embeddings support multilingual CEFR classification without task-specific fine-tuning, and that supervision volume and linguistic grouping explain substantially more variance than model configuration or pretraining proportion.

Assessing whether CEFR proficiency corresponds to a recoverable latent geometric axis in embedding space that generalizes across languages and task modalities (*RQ2*), we demonstrated that CEFR corresponds to a coherent linear direction in embedding space: proficiency is not merely predictable, but geometrically organized as a continuous gradient across languages and task modalities.

These findings shift the perspective from purely predictive evaluation toward representational validity in proficiency assessment. Rather than asking only whether LLMs can classify CEFR levels, we show that they internally encode structured information aligned with standardized proficiency scales. Future work should examine how this geometric structure transfers across languages, interacts with fine-tuning, and relates to interpretable linguistic features.

Limitations

First, embeddings are evaluated in a frozen setting. Fine-tuning may alter the relative impact of pretraining exposure and architectural differences.

Second, language family is used as a coarse proxy for linguistic structure. It does not capture detailed typological variables such as morphological complexity or syntactic configuration.

Third, pretraining proportion is treated as a scalar measure and does not account for domain similarity or data quality.

Fourth, CEFR labels themselves may reflect corpus-specific annotation practices, which can influence performance independently of linguistic competence.

Finally, the recovered proficiency axis may capture linguistic properties that are strongly correlated with CEFR rather than proficiency as an abstract construct. Features such as lexical sophistication, syntactic depth, and discourse complexity are associated with CEFR levels and may serve as proxy variables. While this does not undermine the finding that CEFR is recoverable as a structured representational dimension, it implies that the learned direction may partially reflect correlated linguistic properties associated with CEFR.

Plain Summary

This paper looks at whether large language models (LLMs) understand language proficiency in a meaningful way, rather than just guessing it from surface patterns. Language proficiency is often described using levels like A1 (beginner) to C2 (advanced). In this paper, we ask: do these levels actually exist inside the model's internal representations? In other words, is language proficiency something the model "organises" naturally, or is it just something

we can predict with a classifier? To investigate this, the study uses data covering many languages and proficiency levels. It analyses the internal representations (embeddings) of a multilingual model without fine-tuning it. The findings show two main things:

- The model's representations already contain enough information to predict proficiency levels quite well.
- Proficiency appears to form a continuous scale inside the model: texts from beginner to advanced levels are arranged along a consistent direction, rather than forming unrelated groups.

This suggests that the model does not just memorise patterns, but organises language proficiency in a structured way. The paper also finds that performance differences are mainly explained by factors like dataset size and language differences, rather than by the model itself. Overall, this work shows that language proficiency is not only predictable from LLMs, but is also reflected as an underlying structure in their representations. This could be useful for applications such as adapting text to different learners or generating feedback based on proficiency levels.

Acknowledgements

This study was supported by Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), Brazil.

Bibliographical References

Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. 2016. Automatic text scoring using neural networks. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 715–725.

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater® v. 2. *The Journal of Technology, Learning and Assessment*, 4(3).

Johannes Bjerva, Robert Östling, Maria Han Veiga, Jörg Tiedemann, and Isabelle Augenstein. 2019. What do language representations really represent? *Computational Linguistics*, 45(2):381–389.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Kevyn Collins-Thompson. 2014. Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics*, 165(2):97–135.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8440–8451.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Fei Dong and Yue Zhang. 2016. Automatic features for essay scoring–an empirical study. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1072–1077.

Muhammad Faseeh, Abdul Jaleel, Naeem Iqbal, Anwar Ghani, Akmalbek Abdusalomov, Asif Mehmood, and Young-Im Cho. 2024. Hybrid approach to automated essay scoring: Integrating deep learning embeddings with handcrafted linguistic features for improved accuracy. *Mathematics*, 12(21):3416.

Thomas François and Cédrick Fairon. 2012. An "ai readability" formula for french as a foreign language. In *Proceedings of the 2012 joint conference on empirical methods in Natural Language Processing and computational natural language learning*, pages 466–477.

John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the*

Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4129–4138.

Zixuan Ke and Vincent Ng. 2019. Automated essay scoring: A survey of the state of the art. In *IJCAI*, volume 19, pages 6300–6308.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499.

Xia Li, Huali Yang, Shengze Hu, Jing Geng, Keke Lin, and Yuhai Li. 2022. Enhanced hybrid neural network for automated essay scoring. *Expert Systems*, 39(10):e13068.

Pei Yee Liew and Ian KT Tan. 2024. On automated essay grading using large language models. In *Proceedings of the 2024 8th international conference on computer science and artificial intelligence*, pages 204–211.

Fengkai Liu and John SY Lee. 2023. Hybrid models for sentence readability assessment. In *Proceedings of the 18th workshop on innovative use of nlp for building educational applications (bea 2023)*, pages 448–454.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Improving multi-task deep neural networks via knowledge distillation for natural language understanding. *arXiv preprint arXiv:1904.09482*.

Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M Alves, José Pombal, Nicolas Boizard, et al. 2025a. Eurollm-9b: Technical report. *arXiv preprint arXiv:2506.04079*.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M Guerreiro, Ricardo Rei, Duarte M Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, et al. 2025b. Eurollm: Multilingual language models for europe. *Procedia Computer Science*, 255:53–62.

Elijah Mayfield and Alan W Black. 2020. Should you fine-tune bert for automated essay scoring? In *Proceedings of the fifteenth workshop on innovative use of NLP for building educational applications*, pages 151–162.

Atsushi Mizumoto and Masaki Eguchi. 2023. Exploring the potential of using an ai language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2):100050.

Jesse Mu and Jacob Andreas. 2020. Compositional explanations of neurons. *Advances in Neural Information Processing Systems*, 33:17153–17163.

Council of Europe. 2020. *Common European framework of reference for languages: Companion volume*. Council of Europe.

Council of Europe. Council for Cultural Cooperation. Education Committee. Modern Languages Division. 2001. *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Ellis B Page. 1966. The imminence of... grading essays by computer. *The Phi Delta Kappan*, 47(5):238–243.

Isaac Persing, Alan Davis, and Vincent Ng. 2010. Modeling organization in student essays. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 229–239.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622.

Pedro Uria Rodriguez, Amir Jafari, and Christopher M Ormerod. 2019. Language models and automated essay scoring. *arXiv preprint arXiv:1909.09482*.

Kathrin Seßler, Maurice Fürstenberg, Babette Bühler, and Enkelejda Kasneci. 2025. Can ai grade your essays? a comparative analysis of large language models and teacher ratings in multidimensional essay scoring. In *Proceedings of the 15th international learning analytics and knowledge conference*, pages 462–472.

Mark D Shermis and Jill C Burstein. 2003. *Automated essay scoring: A cross-disciplinary perspective*. Routledge.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1882–1891.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 4593–4601.

Sowmya Vajjala. 2022. Trends, limitations and open challenges in automatic readability assessment research. In *Proceedings of the thirteenth language resources and evaluation conference*, pages 5366–5377.

Sowmya Vajjala and Taraka Rama. 2018. Experiments with universal cefr classification. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 147–153.

Rodrigo Wilkens, Patrick Watrin, Rémi Cardon, Alice Pintard, Isabelle Gribomont, and Thomas François. 2024. Exploring hybrid approaches to readability: experiments on the complementarity between linguistic features and transformers. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2316–2331.

T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, and A. M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Kevin P Yancey, Geoffrey Laflair, Anthony Verardi, and Jill Burstein. 2023. Rating short l2 essays on the cefr scale with gpt-4. In *Proceedings of the 18th workshop on innovative use of NLP for building educational applications (BEA 2023)*, pages 576–584.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading esol texts. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 180–189.

Torsten Zesch, Michael Wojatzki, and Dirk Scholten-Akoun. 2015. Task-independent features for automated essay grading. In *Proceedings of the tenth workshop on innovative use of NLP for building educational applications*, pages 224–232.

7. Language Resource References

Joseph Marvin Imperial, Abdullah Barayan, Regina Stodden, Rodrigo Wilkens, Ricardo Muñoz Sánchez, Lingyun Gao, Melissa Torgbi, Dawn Knight, Gail Forey, Reka R Jablonkai, et al. 2025. Universalcefr: Enabling open multilingual research on language proficiency assessment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9714–9766.

A. Additional Geometric Diagnostics

A.1. Cross-Validated Axis Projection

Figure 4 displays cross-validated projections of embeddings onto the learned proficiency axis for all model variants. Each panel shows the distribution of projection values per CEFR level.

This visualization provides a direct view of ordinal alignment in one dimension. Median projection values increase monotonically from A1 to C2, indicating that higher proficiency levels are systematically shifted along the recovered axis. Although adjacent levels partially overlap, the global ordering remains stable across folds and across model sizes.

Importantly, this figure reflects out-of-fold projections and therefore illustrates generalizable structure rather than in-sample fitting.

A.2. Centroid Distance as a Function of Ordinal Gap

Figure 5 examines geometric separation from a complementary perspective. Instead of showing full distributions, it summarizes each CEFR level by its centroid along the supervised axis and measures the distance between level centroids as a function of absolute ordinal difference.

Distances increase approximately linearly with CEFR gap for all models. This indicates that geometric spacing between levels scales proportionally with ordinal distance, reinforcing the interpretation of proficiency as a continuous gradient rather than a collection of arbitrarily separated clusters.

While Figure 4 emphasizes within-level dispersion and overlap, Figure 5 focuses on between-level scaling. Together, they provide complementary evidence of structured ordinal geometry.

B. Learner and reference dispersion

Figure 6 presents a two-dimensional PCA projection of EuroLLM-9B embeddings without using CEFR labels or task supervision. Points are colored according to task category (learner vs. reference).

The first principal component (PC1) reveals a strong separation between learner and reference texts. Reference texts cluster predominantly on the right-hand side of the projection, while learner texts occupy the left region, with limited overlap in the central area. This indicates that task category accounts for a substantial portion of the dominant variance in embedding space.

In contrast, the second principal component (PC2) primarily captures dispersion within each category rather than cross-category separation. The

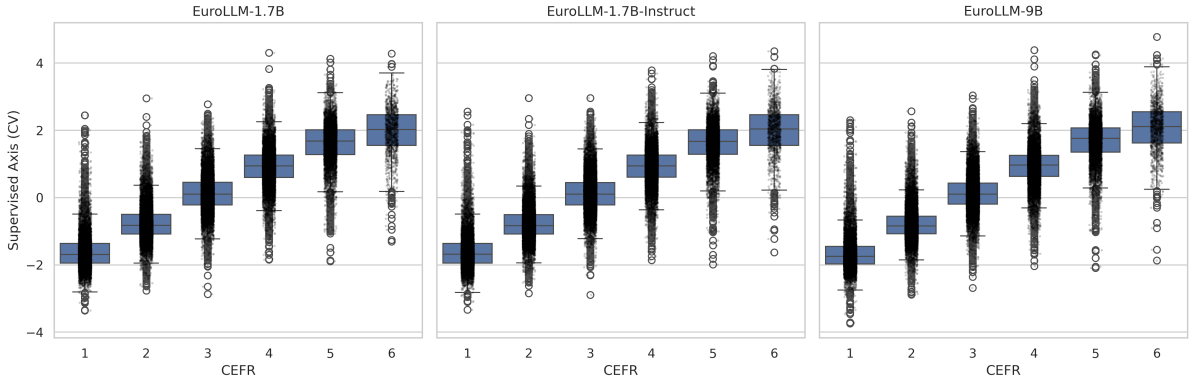


Figure 4: Cross-validated supervised projection onto the CEFR proficiency axis.

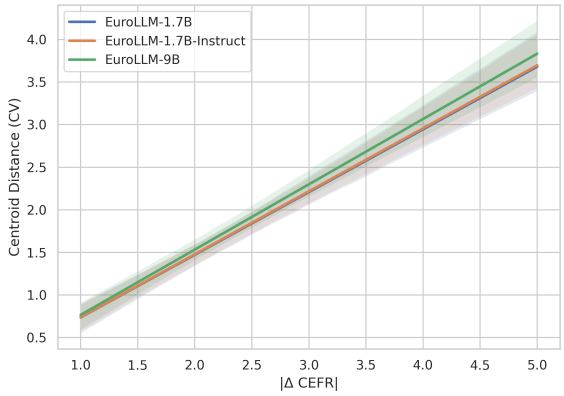


Figure 5: Centroid distance as a function of absolute CEFR gap under cross-validation.

vertical spread suggests substantial internal heterogeneity within both learner and reference groups.

B.1. Performance by Language and Family

Figure 7 shows performance distributions across languages, colored by language family. Romance and Celtic languages tend to exhibit higher average performance, whereas Indo-Aryan languages show comparatively lower scores. However, dispersion within families indicates that language-specific properties also contribute to performance variability, consistent with the mixed-effects analysis.

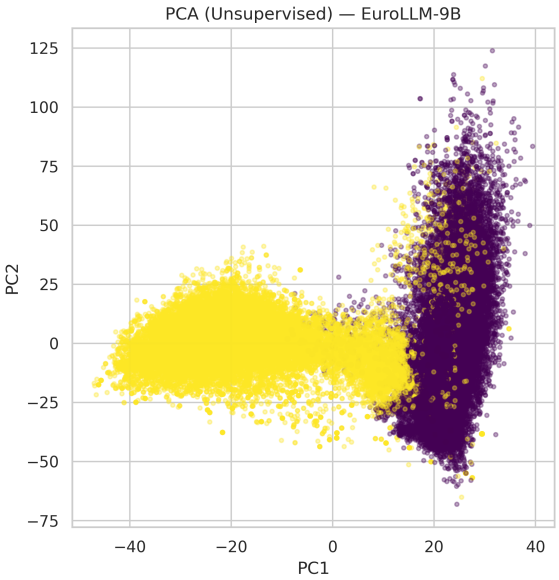


Figure 6: PCA projecting with colors indicating learner and references corpus modalities.

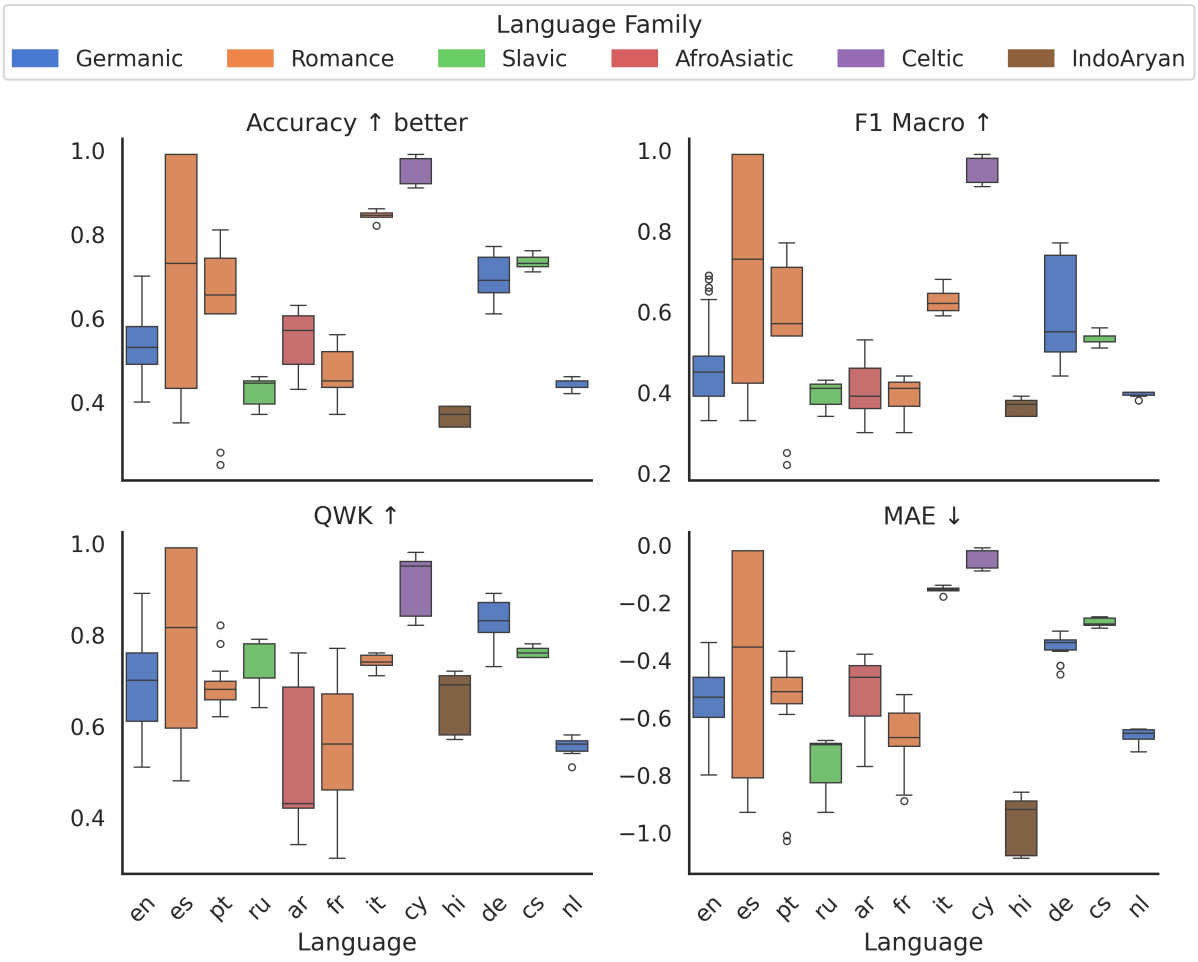


Figure 7: Performance distribution by language (colored by language family) across evaluation metrics.

C. All ML Results

Table 4 presents the complete set of configuration-level results underlying the RQ1 analyses. Each entry corresponds to a unique combination of LLM variant, classifier, language, and corpus. Metrics are reported as mean ± standard deviation across evaluation folds. This detailed table is provided to ensure transparency and reproducibility of all statistical analyses discussed in the main text.

LLMtype	LLMsize	MLmodel	lang	corpus	Acc	F1	QWK	MAE
e	1.700000	LinearSVM	ar	readme	0.43 ± 0.03	0.42 ± 0.03	0.6 ± 0.04	0.77 ± 0.05
e	1.700000	LogReg	ar	readme	0.49 ± 0.02	0.48 ± 0.03	0.73 ± 0.02	0.62 ± 0.02
e	1.700000	MLP	ar	readme	0.51 ± 0.04	0.5 ± 0.04	0.74 ± 0.03	0.59 ± 0.06
e	1.700000	LinearSVM	ar	zaebuc	0.57 ± 0.09	0.41 ± 0.16	0.41 ± 0.16	0.45 ± 0.1
e	1.700000	LogReg	ar	zaebuc	0.62 ± 0.07	0.37 ± 0.09	0.43 ± 0.07	0.39 ± 0.06
e	1.700000	MLP	ar	zaebuc	0.61 ± 0.06	0.35 ± 0.08	0.42 ± 0.11	0.42 ± 0.08
e	1.700000	LinearSVM	cy	learn	0.99 ± 0.01	0.99 ± 0.01	0.98 ± 0.01	0.01 ± 0.01
e	1.700000	LogReg	cy	learn	0.98 ± 0.01	0.98 ± 0.01	0.95 ± 0.02	0.02 ± 0.01
e	1.700000	MLP	cy	learn	0.91 ± 0.02	0.91 ± 0.02	0.82 ± 0.05	0.09 ± 0.02
e	1.700000	LinearSVM	de	elg	0.77 ± 0.06	0.77 ± 0.08	0.89 ± 0.03	0.33 ± 0.08
e	1.700000	LogReg	de	elg	0.76 ± 0.04	0.76 ± 0.05	0.89 ± 0.02	0.33 ± 0.06
e	1.700000	MLP	de	elg	0.73 ± 0.05	0.73 ± 0.05	0.87 ± 0.04	0.37 ± 0.08
e	1.700000	LinearSVM	de	merlin	0.61 ± 0.05	0.45 ± 0.07	0.73 ± 0.05	0.42 ± 0.06
e	1.700000	LogReg	de	merlin	0.67 ± 0.03	0.5 ± 0.06	0.81 ± 0.03	0.34 ± 0.03
e	1.700000	MLP	de	merlin	0.7 ± 0.03	0.56 ± 0.08	0.84 ± 0.02	0.3 ± 0.03
e	1.700000	LinearSVM	en	cambridge	0.62 ± 0.04	0.6 ± 0.04	0.82 ± 0.04	0.49 ± 0.05
e	1.700000	LogReg	en	cambridge	0.68 ± 0.05	0.68 ± 0.05	0.87 ± 0.03	0.39 ± 0.07
e	1.700000	MLP	en	cambridge	0.64 ± 0.05	0.63 ± 0.05	0.89 ± 0.03	0.4 ± 0.07
e	1.700000	LinearSVM	en	cefr	0.44 ± 0.05	0.33 ± 0.06	0.61 ± 0.08	0.66 ± 0.07
e	1.700000	LogReg	en	cefr	0.51 ± 0.04	0.37 ± 0.05	0.7 ± 0.04	0.55 ± 0.05
e	1.700000	MLP	en	cefr	0.47 ± 0.04	0.35 ± 0.06	0.68 ± 0.05	0.6 ± 0.05
e	1.700000	LinearSVM	en	icle500	0.51 ± 0.03	0.44 ± 0.05	0.54 ± 0.08	0.6 ± 0.05
e	1.700000	LogReg	en	icle500	0.55 ± 0.02	0.47 ± 0.06	0.59 ± 0.05	0.53 ± 0.03
e	1.700000	MLP	en	icle500	0.55 ± 0.06	0.44 ± 0.09	0.62 ± 0.07	0.53 ± 0.09
e	1.700000	LinearSVM	en	readme	0.4 ± 0.03	0.35 ± 0.04	0.56 ± 0.05	0.8 ± 0.06
e	1.700000	LogReg	en	readme	0.46 ± 0.03	0.4 ± 0.04	0.7 ± 0.03	0.64 ± 0.04
e	1.700000	MLP	en	readme	0.49 ± 0.02	0.42 ± 0.03	0.73 ± 0.02	0.59 ± 0.03
e	1.700000	LinearSVM	es	caes	0.99 ± 0.0	0.99 ± 0.0	0.99 ± 0.0	0.02 ± 0.0
e	1.700000	LogReg	es	caes	0.99 ± 0.0	0.99 ± 0.0	0.99 ± 0.0	0.02 ± 0.0
e	1.700000	MLP	es	caes	0.99 ± 0.0	0.99 ± 0.0	0.99 ± 0.0	0.02 ± 0.0
e	1.700000	LinearSVM	es	kwiqiz	0.39 ± 0.04	0.36 ± 0.04	0.57 ± 0.07	0.85 ± 0.07
e	1.700000	LogReg	es	kwiqiz	0.44 ± 0.05	0.45 ± 0.05	0.61 ± 0.05	0.73 ± 0.07
e	1.700000	MLP	es	kwiqiz	0.38 ± 0.08	0.38 ± 0.08	0.48 ± 0.08	0.93 ± 0.11
e	1.700000	LinearSVM	fr	kwiqiz	0.55 ± 0.05	0.42 ± 0.05	0.47 ± 0.1	0.57 ± 0.05
e	1.700000	LogReg	fr	kwiqiz	0.56 ± 0.04	0.44 ± 0.06	0.56 ± 0.11	0.52 ± 0.06
e	1.700000	MLP	fr	kwiqiz	0.44 ± 0.03	0.34 ± 0.08	0.31 ± 0.05	0.71 ± 0.02
e	1.700000	LinearSVM	fr	readme	0.37 ± 0.04	0.36 ± 0.04	0.59 ± 0.05	0.89 ± 0.07
e	1.700000	LogReg	fr	readme	0.44 ± 0.04	0.43 ± 0.04	0.74 ± 0.04	0.68 ± 0.05
e	1.700000	MLP	fr	readme	0.43 ± 0.04	0.42 ± 0.05	0.76 ± 0.02	0.67 ± 0.04
e	1.700000	LinearSVM	hi	readme	0.34 ± 0.04	0.34 ± 0.04	0.58 ± 0.03	1.08 ± 0.07
e	1.700000	LogReg	hi	readme	0.36 ± 0.04	0.36 ± 0.04	0.68 ± 0.03	0.94 ± 0.07
e	1.700000	MLP	hi	readme	0.38 ± 0.03	0.38 ± 0.04	0.71 ± 0.03	0.89 ± 0.06
e	1.700000	LinearSVM	it	merlin	0.82 ± 0.04	0.59 ± 0.09	0.71 ± 0.07	0.18 ± 0.04

Continued on next page

LLMtype	LLMsize	MLmodel	lang	corpus	Acc	F1	QWK	MAE
e	1.700000	LogReg	it	merlin	0.85 ± 0.04	0.65 ± 0.1	0.76 ± 0.06	0.15 ± 0.04
e	1.700000	MLP	it	merlin	0.85 ± 0.03	0.61 ± 0.11	0.74 ± 0.06	0.15 ± 0.03
e	1.700000	LinearSVM	nl	elg	0.45 ± 0.02	0.4 ± 0.03	0.56 ± 0.02	0.66 ± 0.02
e	1.700000	LogReg	nl	elg	0.46 ± 0.02	0.4 ± 0.04	0.58 ± 0.03	0.64 ± 0.03
e	1.700000	MLP	nl	elg	0.45 ± 0.03	0.39 ± 0.03	0.56 ± 0.04	0.65 ± 0.04
e	1.700000	LinearSVM	ru	readme	0.37 ± 0.03	0.34 ± 0.03	0.64 ± 0.04	0.93 ± 0.07
e	1.700000	LogReg	ru	readme	0.45 ± 0.04	0.42 ± 0.04	0.78 ± 0.03	0.69 ± 0.06
e	1.700000	MLP	ru	readme	0.45 ± 0.03	0.42 ± 0.04	0.78 ± 0.02	0.69 ± 0.04
e	9.000000	LinearSVM	ar	readme	0.43 ± 0.02	0.44 ± 0.01	0.64 ± 0.03	0.73 ± 0.03
e	9.000000	LogReg	ar	readme	0.49 ± 0.02	0.49 ± 0.04	0.74 ± 0.02	0.6 ± 0.03
e	9.000000	MLP	ar	readme	0.53 ± 0.04	0.53 ± 0.05	0.76 ± 0.03	0.56 ± 0.06
e	9.000000	LinearSVM	ar	zaebuc	0.49 ± 0.08	0.36 ± 0.09	0.42 ± 0.12	0.56 ± 0.1
e	9.000000	LogReg	ar	zaebuc	0.63 ± 0.05	0.34 ± 0.07	0.44 ± 0.1	0.38 ± 0.05
e	9.000000	MLP	ar	zaebuc	0.6 ± 0.06	0.36 ± 0.07	0.36 ± 0.1	0.44 ± 0.07
e	9.000000	LinearSVM	cs	merlin	0.71 ± 0.05	0.52 ± 0.04	0.75 ± 0.04	0.29 ± 0.05
e	9.000000	LogReg	cs	merlin	0.73 ± 0.04	0.54 ± 0.04	0.75 ± 0.06	0.27 ± 0.05
e	9.000000	MLP	cs	merlin	0.75 ± 0.02	0.51 ± 0.07	0.77 ± 0.04	0.25 ± 0.03
e	9.000000	LinearSVM	cy	learn	0.98 ± 0.01	0.98 ± 0.01	0.96 ± 0.02	0.02 ± 0.01
e	9.000000	LogReg	cy	learn	0.97 ± 0.01	0.97 ± 0.01	0.93 ± 0.02	0.03 ± 0.01
e	9.000000	MLP	cy	learn	0.92 ± 0.01	0.92 ± 0.01	0.84 ± 0.03	0.08 ± 0.01
e	9.000000	LinearSVM	de	merlin	0.65 ± 0.04	0.51 ± 0.06	0.78 ± 0.03	0.36 ± 0.04
e	9.000000	LogReg	de	merlin	0.69 ± 0.03	0.54 ± 0.07	0.82 ± 0.02	0.32 ± 0.04
e	9.000000	MLP	de	merlin	0.69 ± 0.01	0.55 ± 0.06	0.83 ± 0.02	0.32 ± 0.02
e	9.000000	LinearSVM	en	cefr	0.49 ± 0.01	0.39 ± 0.01	0.59 ± 0.01	0.64 ± 0.01
e	9.000000	LinearSVM	en	cefr	0.46 ± 0.03	0.37 ± 0.05	0.66 ± 0.05	0.65 ± 0.03
e	9.000000	LogReg	en	cefr	0.57 ± 0.01	0.48 ± 0.03	0.75 ± 0.01	0.46 ± 0.01
e	9.000000	LogReg	en	cefr	0.52 ± 0.04	0.4 ± 0.04	0.71 ± 0.03	0.53 ± 0.03
e	9.000000	MLP	en	cefr	0.57 ± 0.02	0.49 ± 0.03	0.75 ± 0.02	0.46 ± 0.03
e	9.000000	MLP	en	cefr	0.49 ± 0.06	0.35 ± 0.03	0.7 ± 0.04	0.55 ± 0.06
e	9.000000	LinearSVM	en	elg	0.66 ± 0.06	0.66 ± 0.1	0.85 ± 0.04	0.39 ± 0.08
e	9.000000	LogReg	en	elg	0.7 ± 0.05	0.69 ± 0.08	0.88 ± 0.03	0.34 ± 0.06
e	9.000000	MLP	en	elg	0.65 ± 0.05	0.63 ± 0.08	0.86 ± 0.04	0.39 ± 0.07
e	9.000000	LinearSVM	en	icle500	0.53 ± 0.06	0.48 ± 0.08	0.6 ± 0.07	0.55 ± 0.08
e	9.000000	LogReg	en	icle500	0.54 ± 0.06	0.48 ± 0.08	0.61 ± 0.05	0.53 ± 0.06
e	9.000000	MLP	en	icle500	0.57 ± 0.05	0.47 ± 0.08	0.61 ± 0.03	0.52 ± 0.05
e	9.000000	LinearSVM	en	readme	0.41 ± 0.04	0.36 ± 0.05	0.6 ± 0.03	0.76 ± 0.05
e	9.000000	LogReg	en	readme	0.48 ± 0.02	0.42 ± 0.04	0.72 ± 0.03	0.61 ± 0.04
e	9.000000	MLP	en	readme	0.5 ± 0.02	0.44 ± 0.03	0.75 ± 0.02	0.57 ± 0.03
e	9.000000	LinearSVM	es	caes	0.99 ± 0.0	0.99 ± 0.0	0.99 ± 0.0	0.02 ± 0.0
e	9.000000	LogReg	es	caes	0.99 ± 0.0	0.99 ± 0.0	0.99 ± 0.0	0.02 ± 0.0
e	9.000000	MLP	es	caes	0.99 ± 0.0	0.99 ± 0.0	0.99 ± 0.0	0.02 ± 0.0
e	9.000000	LinearSVM	es	kwiqiz	0.43 ± 0.07	0.42 ± 0.07	0.59 ± 0.08	0.82 ± 0.1
e	9.000000	LogReg	es	kwiqiz	0.44 ± 0.05	0.44 ± 0.06	0.61 ± 0.07	0.76 ± 0.1
e	9.000000	MLP	es	kwiqiz	0.44 ± 0.05	0.43 ± 0.05	0.61 ± 0.04	0.78 ± 0.07
e	9.000000	LinearSVM	fr	kwiqiz	0.48 ± 0.03	0.39 ± 0.06	0.44 ± 0.11	0.69 ± 0.07
e	9.000000	LogReg	fr	kwiqiz	0.52 ± 0.04	0.41 ± 0.06	0.52 ± 0.12	0.57 ± 0.08
e	9.000000	MLP	fr	kwiqiz	0.47 ± 0.05	0.33 ± 0.06	0.46 ± 0.09	0.64 ± 0.08
e	9.000000	LinearSVM	hi	readme	0.34 ± 0.03	0.34 ± 0.03	0.57 ± 0.06	1.09 ± 0.08
e	9.000000	LogReg	hi	readme	0.39 ± 0.03	0.39 ± 0.03	0.71 ± 0.04	0.88 ± 0.06
e	9.000000	MLP	hi	readme	0.39 ± 0.05	0.39 ± 0.05	0.72 ± 0.03	0.86 ± 0.06
e	9.000000	LinearSVM	it	merlin	0.84 ± 0.02	0.6 ± 0.11	0.74 ± 0.04	0.16 ± 0.02
e	9.000000	LogReg	it	merlin	0.86 ± 0.03	0.68 ± 0.14	0.76 ± 0.05	0.14 ± 0.03
e	9.000000	MLP	it	merlin	0.84 ± 0.03	0.63 ± 0.12	0.73 ± 0.04	0.16 ± 0.03
e	9.000000	LinearSVM	nl	elg	0.42 ± 0.02	0.38 ± 0.03	0.51 ± 0.02	0.72 ± 0.02
e	9.000000	LogReg	nl	elg	0.45 ± 0.02	0.4 ± 0.03	0.57 ± 0.03	0.64 ± 0.03
e	9.000000	MLP	nl	elg	0.43 ± 0.03	0.4 ± 0.04	0.54 ± 0.03	0.68 ± 0.04
e	9.000000	LinearSVM	pt	cople2	0.75 ± 0.05	0.71 ± 0.06	0.72 ± 0.06	0.52 ± 0.09
e	9.000000	LogReg	pt	cople2	0.81 ± 0.05	0.77 ± 0.08	0.82 ± 0.04	0.37 ± 0.08
e	9.000000	MLP	pt	cople2	0.25 ± 0.05	0.22 ± 0.05	0.62 ± 0.03	1.03 ± 0.09
e	9.000000	LinearSVM	pt	peapl2	0.66 ± 0.06	0.6 ± 0.07	0.69 ± 0.07	0.46 ± 0.09
e	9.000000	LogReg	pt	peapl2	0.66 ± 0.04	0.58 ± 0.06	0.69 ± 0.09	0.46 ± 0.07
e	9.000000	MLP	pt	peapl2	0.65 ± 0.05	0.56 ± 0.07	0.68 ± 0.09	0.47 ± 0.07
e	9.000000	LinearSVM	ru	readme	0.38 ± 0.04	0.36 ± 0.04	0.68 ± 0.03	0.87 ± 0.06
e	9.000000	LogReg	ru	readme	0.46 ± 0.04	0.43 ± 0.05	0.79 ± 0.02	0.68 ± 0.05
e	9.000000	MLP	ru	readme	0.44 ± 0.03	0.4 ± 0.03	0.78 ± 0.02	0.7 ± 0.05
ei	1.700000	LinearSVM	ar	zaebuc	0.6 ± 0.09	0.39 ± 0.1	0.43 ± 0.16	0.42 ± 0.11
ei	1.700000	LogReg	ar	zaebuc	0.62 ± 0.05	0.37 ± 0.08	0.42 ± 0.08	0.39 ± 0.05
ei	1.700000	MLP	ar	zaebuc	0.57 ± 0.05	0.3 ± 0.02	0.34 ± 0.12	0.46 ± 0.09
ei	1.700000	LinearSVM	cs	merlin	0.73 ± 0.03	0.54 ± 0.03	0.75 ± 0.03	0.28 ± 0.03
ei	1.700000	LogReg	cs	merlin	0.76 ± 0.02	0.56 ± 0.02	0.78 ± 0.02	0.25 ± 0.03
ei	1.700000	MLP	cs	merlin	0.72 ± 0.03	0.54 ± 0.08	0.77 ± 0.03	0.28 ± 0.03
ei	1.700000	LinearSVM	cy	learn	0.99 ± 0.01	0.99 ± 0.01	0.98 ± 0.01	0.01 ± 0.01
ei	1.700000	LogReg	cy	learn	0.98 ± 0.01	0.98 ± 0.01	0.95 ± 0.03	0.02 ± 0.01
ei	1.700000	MLP	cy	learn	0.91 ± 0.02	0.91 ± 0.02	0.82 ± 0.03	0.09 ± 0.02
ei	1.700000	LinearSVM	de	elg	0.76 ± 0.07	0.76 ± 0.08	0.87 ± 0.04	0.35 ± 0.09
ei	1.700000	LogReg	de	elg	0.76 ± 0.04	0.75 ± 0.05	0.89 ± 0.03	0.33 ± 0.06
ei	1.700000	MLP	de	elg	0.69 ± 0.05	0.67 ± 0.07	0.84 ± 0.04	0.45 ± 0.09
ei	1.700000	LinearSVM	de	merlin	0.61 ± 0.03	0.44 ± 0.06	0.74 ± 0.05	0.42 ± 0.05
ei	1.700000	LogReg	de	merlin	0.66 ± 0.04	0.49 ± 0.06	0.8 ± 0.03	0.35 ± 0.04
ei	1.700000	MLP	de	merlin	0.66 ± 0.04	0.5 ± 0.04	0.81 ± 0.03	0.34 ± 0.05
ei	1.700000	LinearSVM	en	cambridge	0.63 ± 0.02	0.61 ± 0.02	0.82 ± 0.04	0.48 ± 0.04
ei	1.700000	LogReg	en	cambridge	0.69 ± 0.05	0.69 ± 0.06	0.87 ± 0.03	0.37 ± 0.07
ei	1.700000	MLP	en	cambridge	0.62 ± 0.03	0.61 ± 0.04	0.85 ± 0.02	0.46 ± 0.03

Continued on next page

LLMtype	LLMsize	MLmodel	lang	corpus	Acc	F1	QWK	MAE
ei	1.700000	LinearSVM	en	cefr	0.53 ± 0.01	0.44 ± 0.02	0.68 ± 0.01	0.55 ± 0.02
ei	1.700000	LogReg	en	cefr	0.57 ± 0.01	0.48 ± 0.02	0.75 ± 0.01	0.47 ± 0.01
ei	1.700000	MLP	en	cefr	0.58 ± 0.01	0.49 ± 0.02	0.76 ± 0.01	0.44 ± 0.01
ei	1.700000	LinearSVM	en	elg	0.61 ± 0.05	0.63 ± 0.06	0.81 ± 0.04	0.47 ± 0.08
ei	1.700000	LogReg	en	elg	0.66 ± 0.06	0.65 ± 0.07	0.86 ± 0.02	0.38 ± 0.06
ei	1.700000	MLP	en	elg	0.64 ± 0.04	0.59 ± 0.06	0.85 ± 0.02	0.41 ± 0.06
ei	1.700000	LinearSVM	en	icle500	0.49 ± 0.03	0.42 ± 0.05	0.51 ± 0.08	0.63 ± 0.05
ei	1.700000	LogReg	en	icle500	0.52 ± 0.02	0.45 ± 0.05	0.58 ± 0.04	0.58 ± 0.02
ei	1.700000	MLP	en	icle500	0.53 ± 0.07	0.45 ± 0.08	0.58 ± 0.07	0.57 ± 0.08
ei	1.700000	LinearSVM	es	caes	0.99 ± 0.0	0.99 ± 0.0	0.99 ± 0.0	0.02 ± 0.0
ei	1.700000	LogReg	es	caes	0.99 ± 0.0	0.99 ± 0.0	0.99 ± 0.0	0.02 ± 0.0
ei	1.700000	MLP	es	caes	0.99 ± 0.0	0.99 ± 0.0	0.99 ± 0.0	0.02 ± 0.0
ei	1.700000	LinearSVM	es	kwiqiz	0.35 ± 0.07	0.33 ± 0.07	0.52 ± 0.13	0.92 ± 0.12
ei	1.700000	LogReg	es	kwiqiz	0.47 ± 0.06	0.47 ± 0.06	0.64 ± 0.06	0.69 ± 0.09
ei	1.700000	MLP	es	kwiqiz	0.38 ± 0.09	0.38 ± 0.1	0.52 ± 0.05	0.87 ± 0.12
ei	1.700000	LinearSVM	fr	kwiqiz	0.52 ± 0.05	0.4 ± 0.05	0.46 ± 0.11	0.6 ± 0.07
ei	1.700000	LogReg	fr	kwiqiz	0.54 ± 0.05	0.42 ± 0.04	0.56 ± 0.11	0.53 ± 0.06
ei	1.700000	MLP	fr	kwiqiz	0.43 ± 0.04	0.3 ± 0.07	0.35 ± 0.12	0.73 ± 0.07
ei	1.700000	LinearSVM	fr	readme	0.38 ± 0.04	0.37 ± 0.04	0.6 ± 0.06	0.87 ± 0.06
ei	1.700000	LogReg	fr	readme	0.44 ± 0.04	0.43 ± 0.05	0.74 ± 0.03	0.68 ± 0.05
ei	1.700000	MLP	fr	readme	0.45 ± 0.02	0.44 ± 0.02	0.77 ± 0.03	0.65 ± 0.04
ei	1.700000	LinearSVM	hi	readme	0.34 ± 0.03	0.34 ± 0.03	0.58 ± 0.03	1.09 ± 0.04
ei	1.700000	LogReg	hi	readme	0.37 ± 0.05	0.37 ± 0.05	0.69 ± 0.02	0.92 ± 0.06
ei	1.700000	MLP	hi	readme	0.39 ± 0.02	0.38 ± 0.02	0.69 ± 0.02	0.9 ± 0.05
ei	1.700000	LinearSVM	pt	cople2	0.74 ± 0.05	0.71 ± 0.07	0.68 ± 0.08	0.59 ± 0.1
ei	1.700000	LogReg	pt	cople2	0.79 ± 0.04	0.76 ± 0.07	0.78 ± 0.05	0.42 ± 0.08
ei	1.700000	MLP	pt	cople2	0.28 ± 0.03	0.25 ± 0.03	0.62 ± 0.04	1.01 ± 0.06
ei	1.700000	LinearSVM	pt	peapl2	0.61 ± 0.04	0.54 ± 0.04	0.68 ± 0.03	0.5 ± 0.04
ei	1.700000	LogReg	pt	peapl2	0.61 ± 0.03	0.54 ± 0.06	0.66 ± 0.05	0.52 ± 0.04
ei	1.700000	MLP	pt	peapl2	0.61 ± 0.03	0.54 ± 0.05	0.65 ± 0.08	0.54 ± 0.06

Table 4: Results of all model

Author Index

Aissa, Wafa, 12
Alshatti, Abdullah, 193
Alva-Manchego, Fernando, 193
Amaro, Raquel, 12
Antunes, David, 12, 151

Bakker, Jan, 121
Baptista, Jorge, 12, 151
Bian, Kexin, 49
Biemann, Chris, 1
Bigi, Brigitte, 61
Briscoe, Ted, 74

Cardon, Rémi, 164, 210

Dell'Orletta, Felice, 89
Dogruoz, A. Seza, 164, 210
Doueihi, Julien Zakhia, 12
Drevet, Ludivine Javourey, 61

François, Thomas, 12, 61, 181

Gala, Núria, 61
Ganesh, Ananya, 130
Gao, Yingqiang, 26
Garcia, Marcos, 101

Hubarava, Hanna, 26

Kalinina, Anna, 181
Kamps, Jaap, 121
Karumbaiah, Shamya, 130
Komachi, Mamoru, 49

Maheshwary, Pragati, 130
Mamede, Nuno, 151
Mathas, Pascal, 121

Norré, Magali, 61

Ozten, Akchay, 142

Papucci, Michele, 89

Qwaider, Chatrine, 74

Rabih, Nour, 74
Ribeiro, Eugénio, 12, 151
Rodríguez Rey, Sandra, 101

Schockaert, Steven, 193
Schomacker, Thorben, 1

Todirascu, Amalia, 181
Tran Pham, Hanh Trang, 12
Tropmann-Frick, Marina, 1

Vassiliadou, Hélène, 181
Venturi, Giulia, 89

Wilkens, Rodrigo, 142, 227

Yimam, Seid Muhie, 1
Yoon, Su-Youn, 49