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.