

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

Hanna Hubarava^{1,2} Yingqiang Gao^{*1}

¹Department of Computational Linguistics, University of Zurich, Switzerland

²Department 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-EAS_I (Basu et al., 2023); SIMPA (Scarton et al., 2018); WIKILARGE (Zhang and Lapata, 2017); NEWS_{EL}A (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

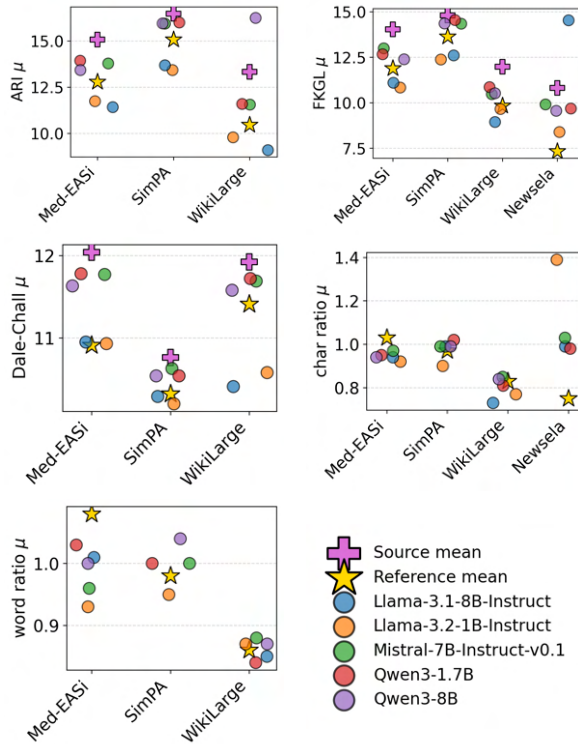


Figure 1: Mean control-attribute values across models and datasets. Fine-tuned models generate text that is simpler than the source but more complex than the target. Sentence-aligned datasets (MED-EASi, SIMPA, WIKILARGE) show almost no compression signal, limiting training and evaluation.

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*.

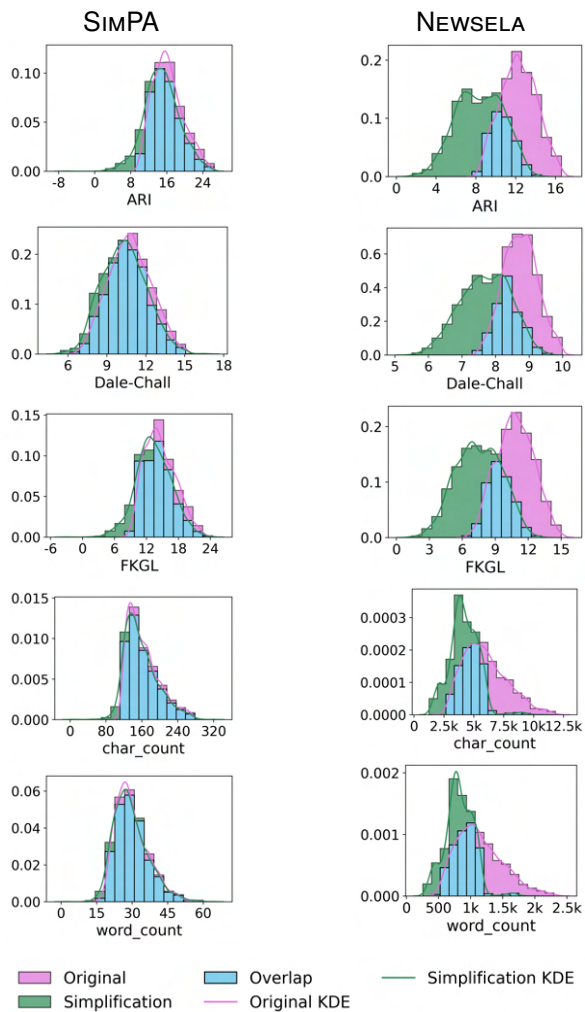


Figure 2: Both SIMPA and NEWSELA show larger distribution shifts by readability than by length transformations. Text-aligned NEWSELA shows greater distribution shift than the sentence-aligned SIMPA.

- 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 ($\langle \text{FKGL}=5 \rangle$). 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 ($\langle \text{WORD_COMPRESSION}=0.5 \rangle$).

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 NEWSLA 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 (Maddala 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.

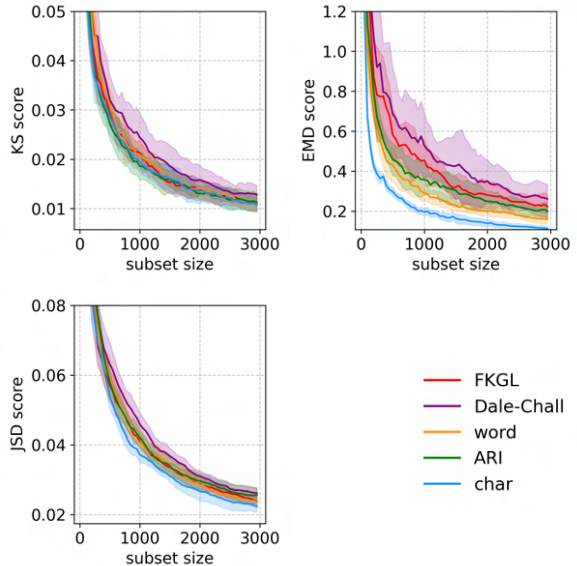


Figure 4: WIKILARGE. *Global* sampling with stratification by readability and length, measured in terms of KS, EMD and JSD. Stratification by N chars shows smallest divergence across all metrics.

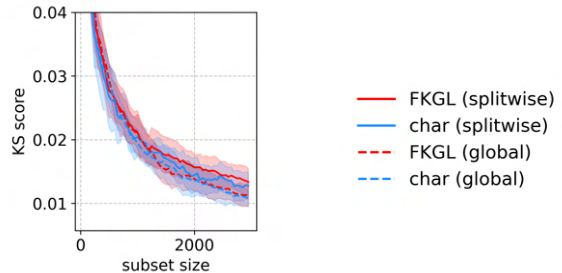


Figure 5: WIKILARGE. A comparison of split-based and split-agnostic sampling shows the latter yields lower divergence across both length and readability, with FKGL and N chars yielding similar results.

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
NEWSLA 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
NEWSLA 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 NEWSLA. 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). NEWSLA (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 \uparrow									(b) LENS \uparrow									(c) MAE \downarrow													
Llama				Qwen					Llama				Qwen					Llama				Qwen									
1B	3B	8B	13B	1.7B	4B	8B	14B	1B	3B	8B	13B	1.7B	4B	8B	14B	1B	3B	8B	13B	1.7B	4B	8B	14B	1B	3B	8B	13B	1.7B	4B	8B	14B
42.64	43.05	43.99	52.02	50.50	51.17	51.10	<u>51.93</u>	57.33	56.82	59.90	57.41	58.18	58.96	<u>59.27</u>	58.88	3.96	3.49	2.91	2.99	2.72	2.91	2.34	<u>2.61</u>	2.56	1.11	2.18	1.27	1.22	1.08	1.03	<u>1.06</u>
42.15	60.16	41.81	67.78	65.52	65.07	<u>67.65</u>	65.90	55.05	59.13	55.42	58.35	58.10	58.09	<u>58.59</u>	58.38																

(d) SARI \uparrow									(e) LENS \uparrow									(f) MAE \downarrow													
Llama				Qwen					Llama				Qwen					Llama				Qwen									
1B	3B	8B	13B	1.7B	4B	8B	14B	1B	3B	8B	13B	1.7B	4B	8B	14B	1B	3B	8B	13B	1.7B	4B	8B	14B	1B	3B	8B	13B	1.7B	4B	8B	14B
42.43	41.58	45.25	<u>52.06</u>	52.09	51.33	51.91	50.72	55.74	58.45	59.83	57.47	60.54	58.91	<u>59.94</u>	59.19	0.39	0.36	0.21	0.29	<u>0.18</u>	0.22	0.19	0.15	0.18	0.04	0.04	0.05	0.07	<u>0.05</u>	0.04	0.04
38.95	<u>64.29</u>	59.14	67.17	62.82	62.30	65.96	62.94	53.93	<u>58.45</u>	57.54	58.28	58.29	57.55	58.49	57.81																

Table 3: Scaling experiment with a broader parameter size range. Rows in green represent results on MED-EAS_I, 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.

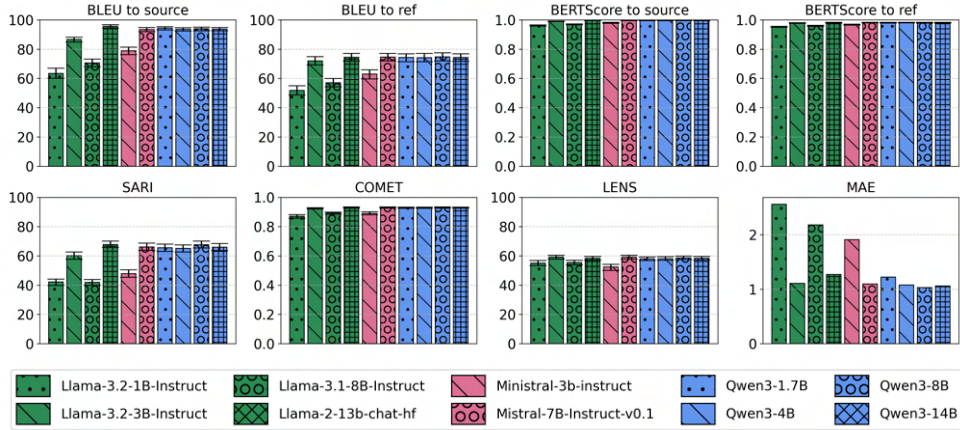


Figure 6: Dataset: SIMPA. Control attribute: FKGL. Scaling does not always boost performance. However, we observe strong positive correlation between SARI and COMET, and a strong negative correlation between SARI/COMET and error-based metrics across most datasets and control attributes.

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-EAS_I 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 \uparrow						LENS \uparrow				MAE \downarrow							
		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 divergences 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 NEWSLA 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 Formula.
- 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, Léo 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.

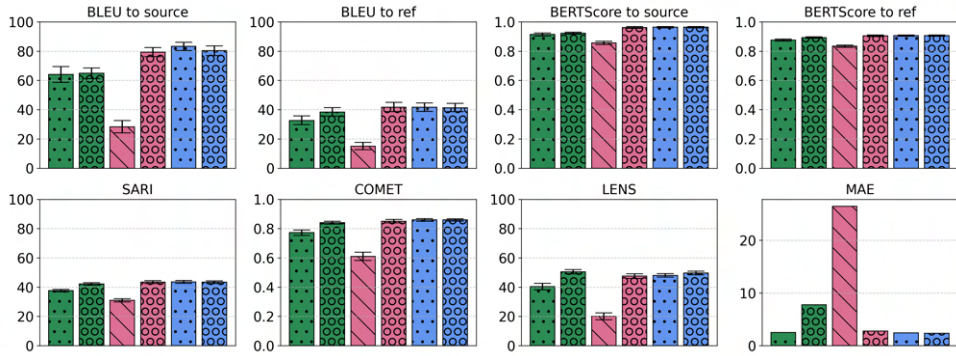
- Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2023. LENS: A learnable evaluation metric for text simplification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16383–16408, Toronto, Canada. Association for Computational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

(a) FKGL



(b) char compression

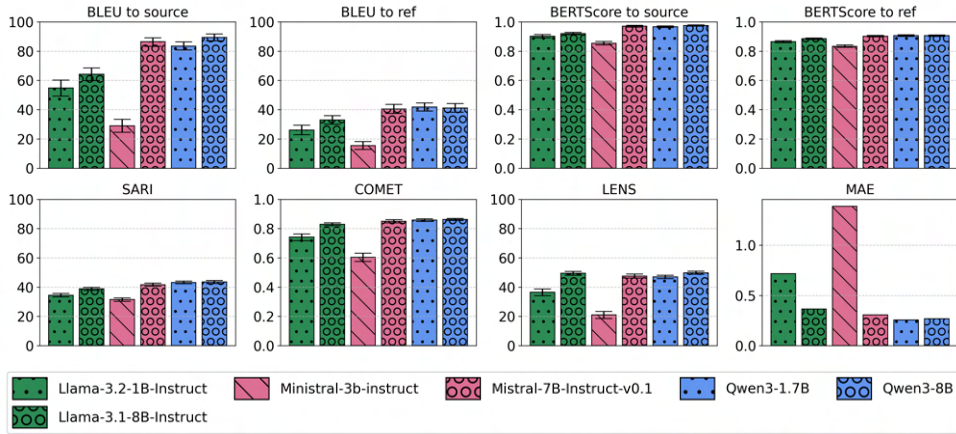
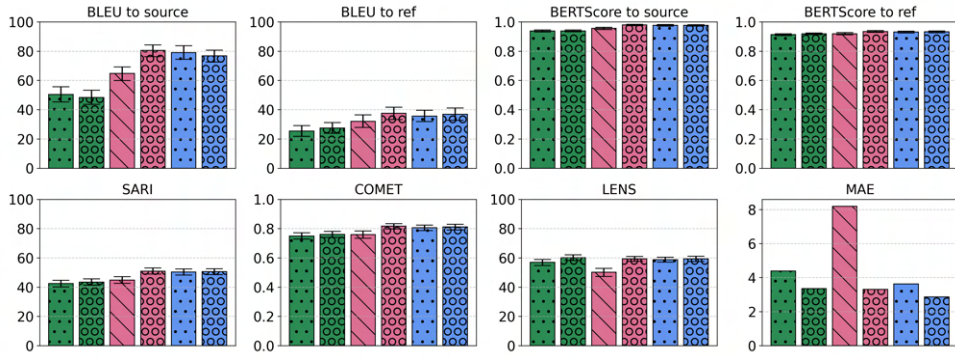
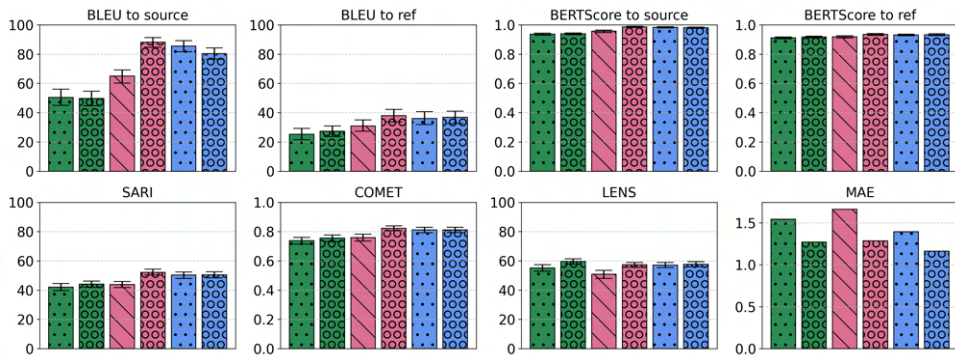


Figure 8: Model performance on MED-EASi.

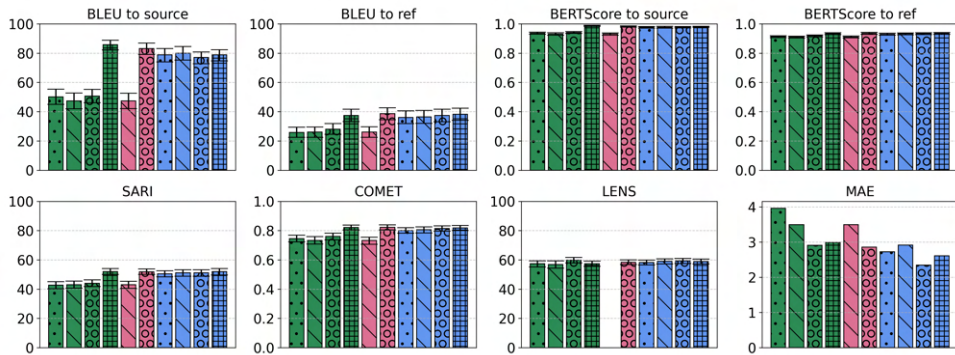
(a) ARI



(b) Dale-Chall



(c) FKGL



(d) char compression

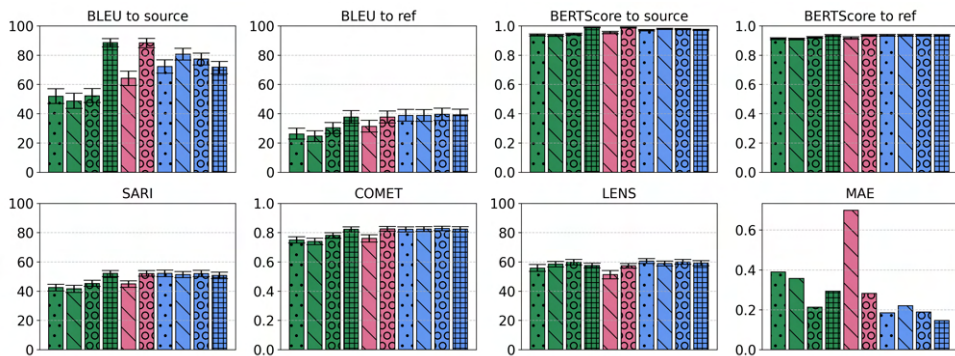
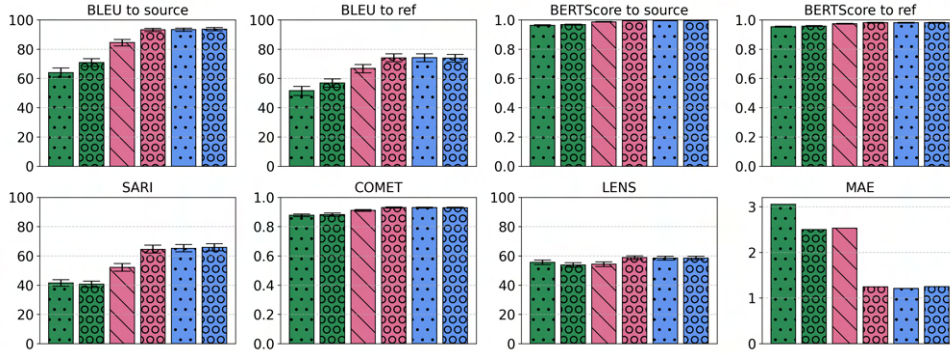
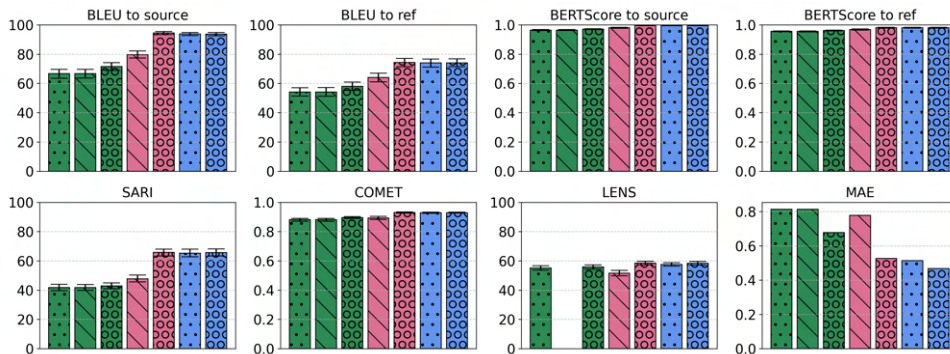


Figure 9: Model performance on SIMPA.

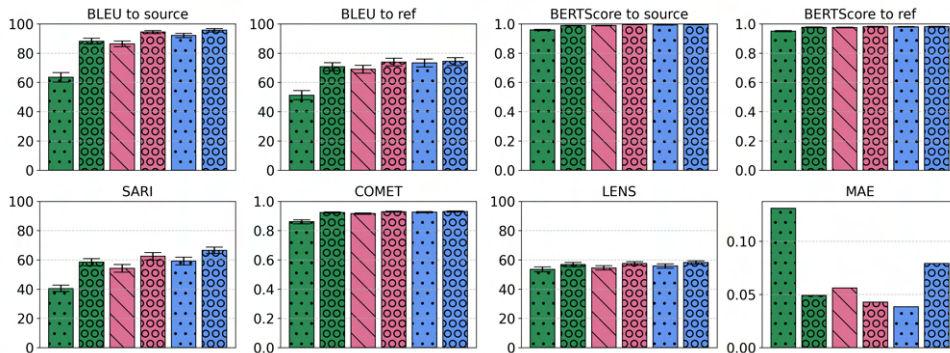
(a) ARI



(b) Dale-Chall



(c) word compression



(d) char compression

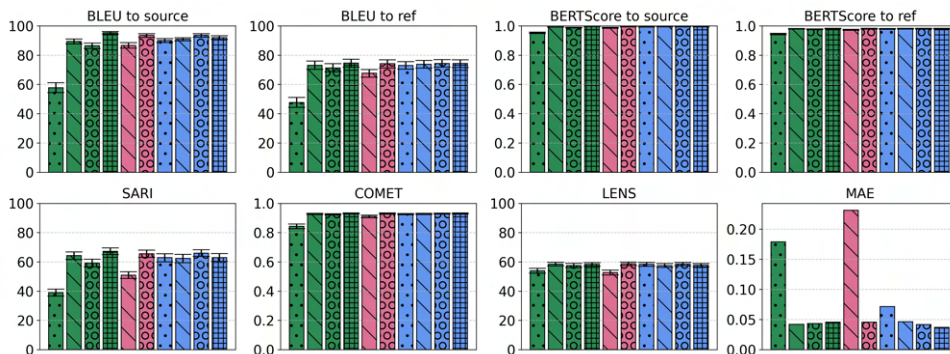
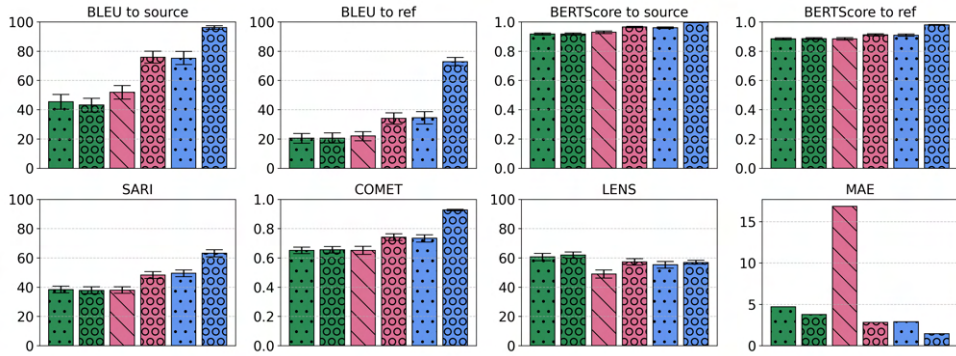
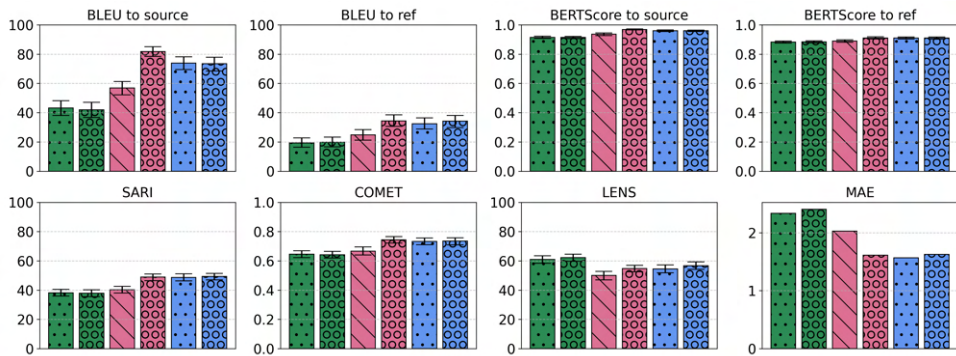


Figure 10: Model performance on WikiLARGE.

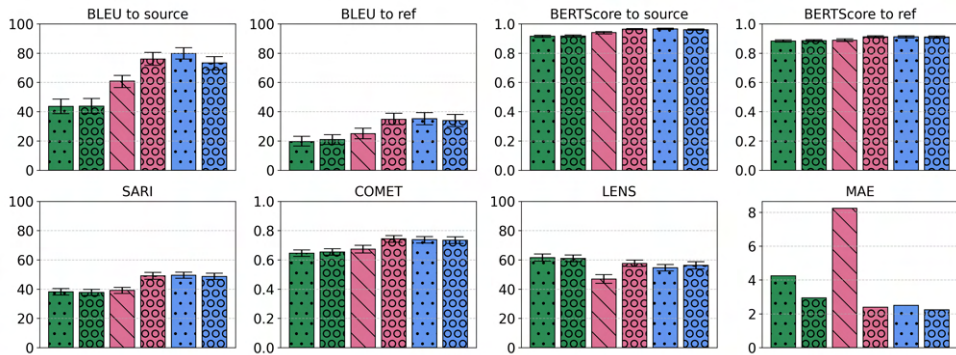
(a) ARI



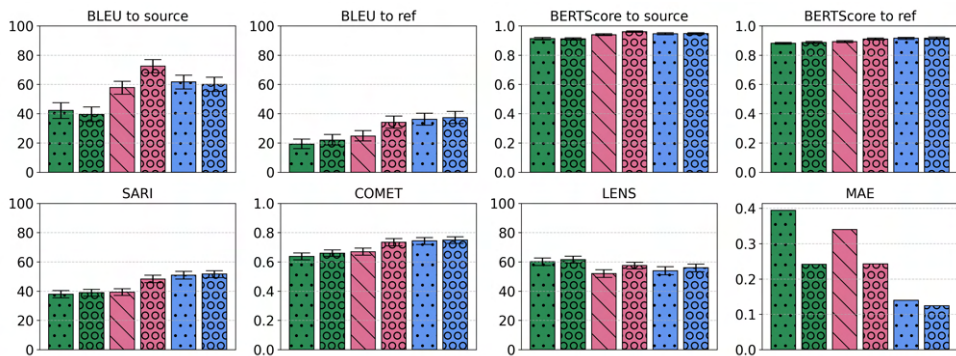
(b) Dale-Chall



(c) FKGL



(d) char compression



B. Baseline Controllability Analysis

Figure 11: Dataset: SIMPA

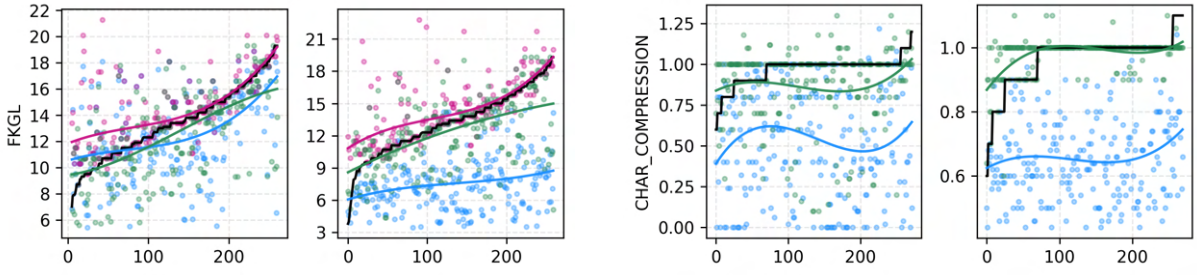


Figure 12: Dataset: MEDEASi

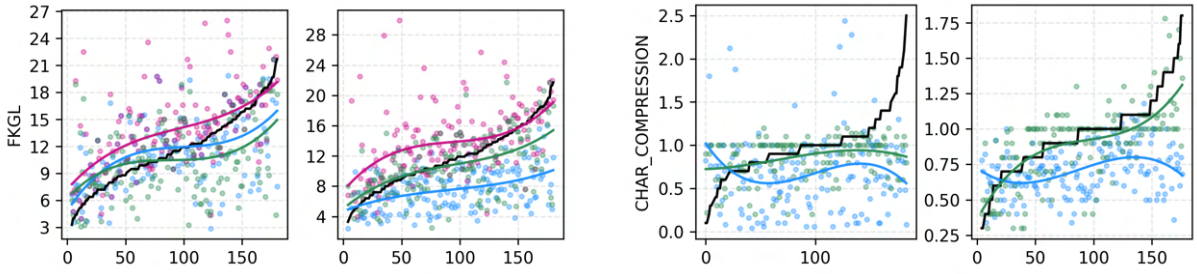


Figure 13: Dataset: WIKILARGE

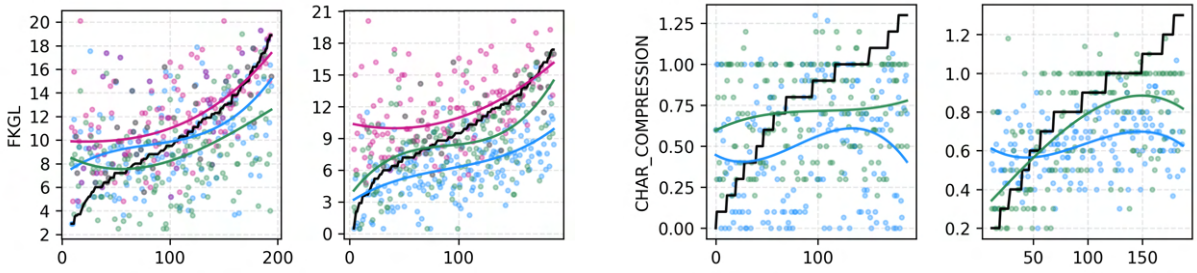
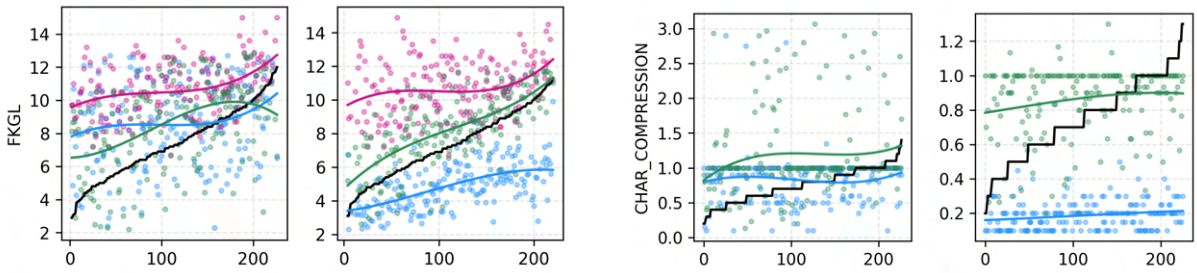


Figure 14: Dataset: NEWSLEA



● Source — Reference ● Baseline ● Finetuned
— Source fit — Baseline fit — Finetuned fit

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: NEWSLA.

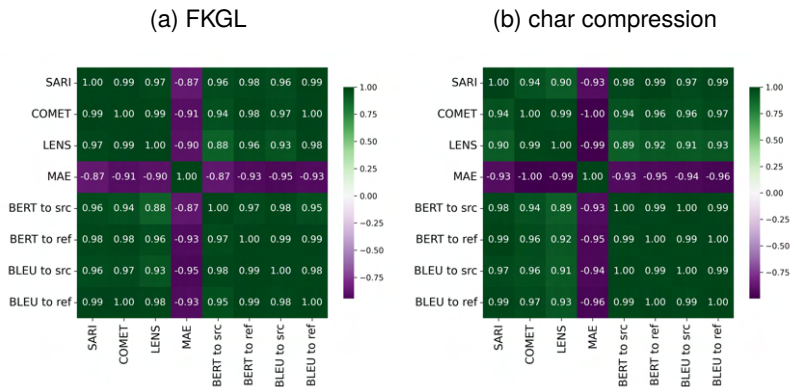


Figure 17: Pearson correlation heatmaps analyzing metric correlation, aggregated across all models. Dataset: MED-EASi.

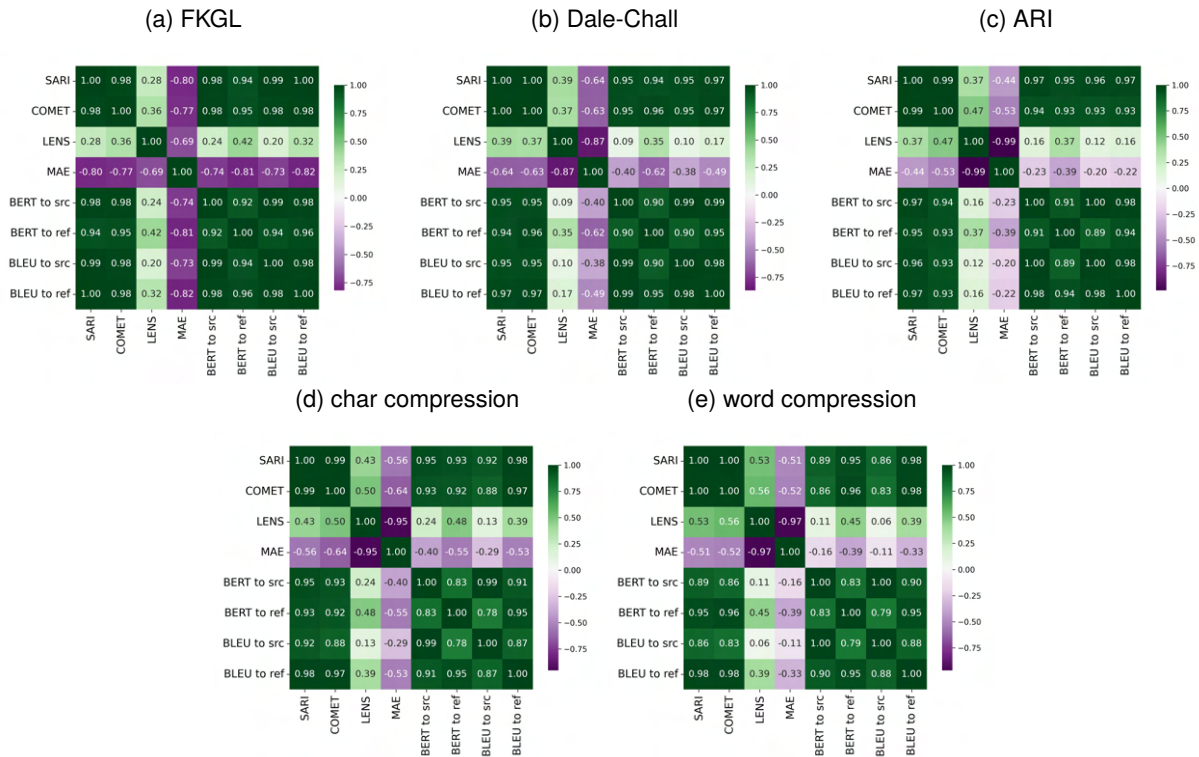


Figure 18: Pearson correlation heatmaps analyzing metric correlation, aggregated across all models. Dataset: SIMPA.

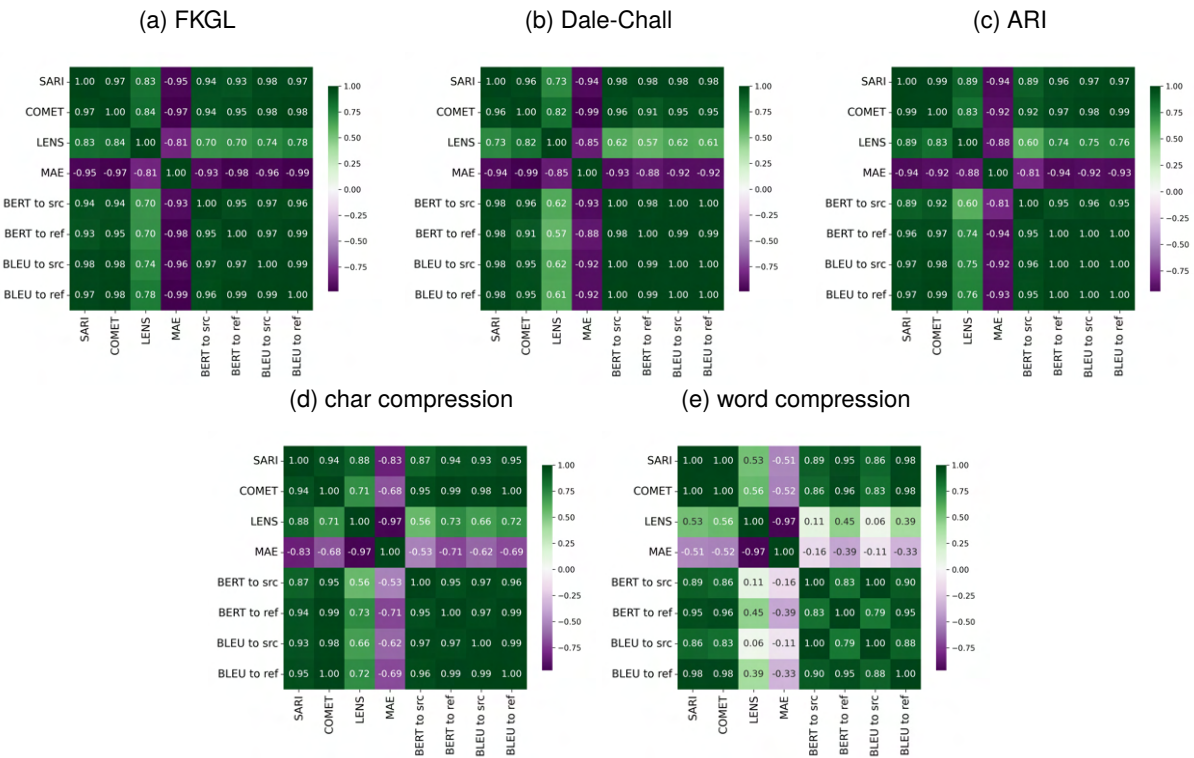
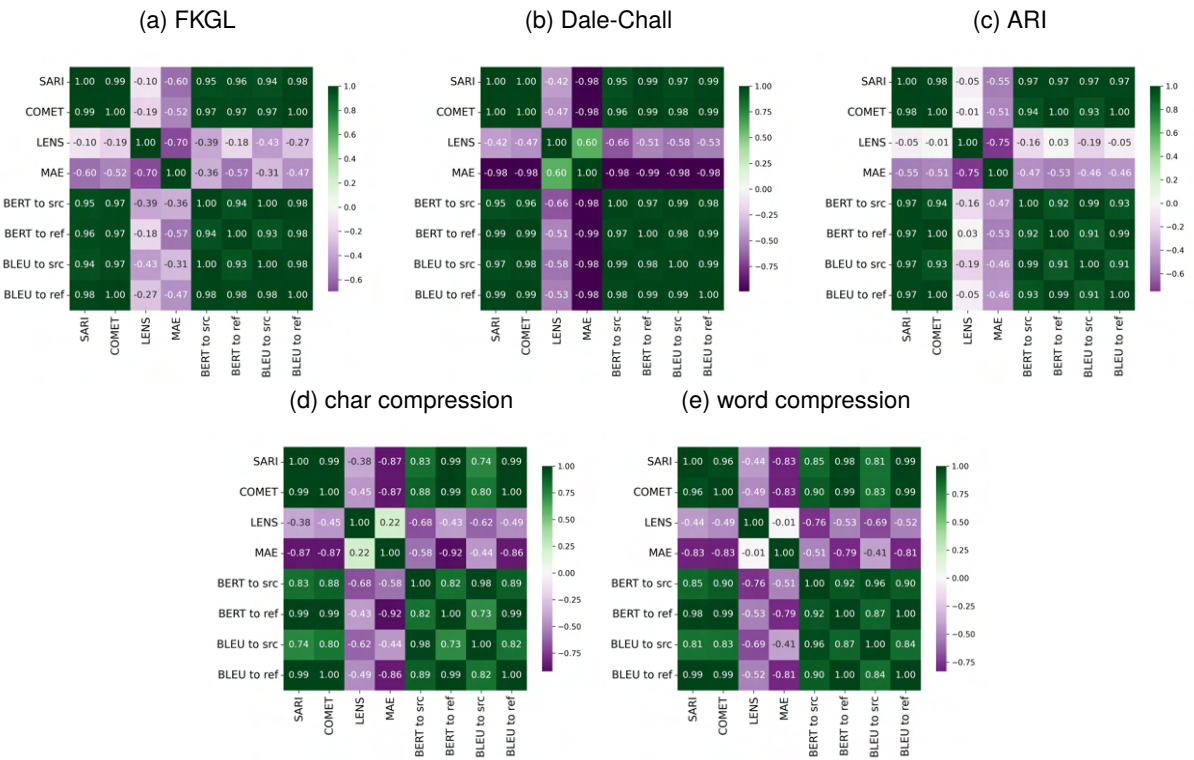


Figure 19: Pearson correlation heatmaps analyzing metric correlation, aggregated across all models. Dataset: WIKILARGE.



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-EAS1. 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	PEFT
<pre> 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 </pre>	<pre> 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 </pre>