

SIMPLE: A Comparative Study of Prompting and Fine-Tuning for LLM-Based Text Simplification

Zixian Meng Dongqiang Yang*

School of Computing and Artificial Intelligence

Shandong Jianzhu University, China

2024110110@stu.sdjzu.edu.cn, ydq@sdjzu.edu.cn

Abstract

Fine-tuning and prompting are two widely used approaches for adapting large language models to downstream tasks. However, their effectiveness on text simplification remains underexplored. We present **SIMPLE**, a comprehensive evaluation of large language model-based sentence simplification. We evaluate 12 open-weight large language models across diverse domains, including Wikipedia, medical, scientific, and news texts. In addition to standard evaluation metrics, we conduct linguistic and targeted manual analyses of the generated simplifications. Our findings show that: 1) Fine-tuning outperforms prompting on Wikipedia-based simplification, whereas prompting performs better in non-Wikipedia domains; 2) fine-tuned models exhibit a strong alignment with human-written references across target readability levels, while prompting shows limited controllability over output difficulty; and 3) the primary discrepancies between LLM-generated simplifications and human references arise from syntactic restructuring and discourse coherence.

1 Introduction

Large language models (LLMs) have substantially advanced text simplification (TS). Existing LLM-based TS research has explored two main adaptation paradigms—prompting and fine-tuning—as well as a range of evaluation frameworks. Prompting-based approaches have demonstrated strong performance across various simplification tasks (Feng et al., 2023; Kew et al., 2023; Qiang et al., 2025; Zhang et al., 2026), while fine-tuning approaches, including parameter-efficient methods such as LoRA and QLoRA, have been applied to lexical and sentence simplification (Baez and Saggion, 2023; Cohen et al., 2026). Concurrently, researchers have developed diverse evaluation frameworks, including LLM-as-a-Judge assessment (Barbu et al., 2025) and NLP-based readability metrics (Crossley, 2025).

Despite these advances, several important questions remain unresolved. First, existing comparisons of fine-tuning and prompting remain limited in scope. For example, Simplify-This (Cohen et al., 2026) focused exclusively on encoder-decoder models (e.g., BART (Lewis et al., 2020) and T5 (Raffel et al., 2020), up to 400M parameters), leaving decoder-only LLMs largely unexplored. Moreover, a substantial portion of recent work evaluates instruction-tuned models, making it difficult to disentangle the effects of adaptation strategies from those of instruction tuning (Itzhak et al., 2024). Second, controllable text simplification across readability levels remains underexplored. Although Newsela has been widely used in recent LLM-based TS research (Kew et al., 2023; Qiang et al., 2025; Cohen et al., 2026), its multiple readability levels have seldom been exploited to systematically evaluate controllable simplification. Third, standard evaluation metrics such as LENS, FKGL, and SARI provide limited insight into the linguistic properties of simplification outputs (Xu et al., 2016; Sulem et al., 2018; Alva-Manchego et al., 2021; Ilgen and Hattab, 2025). As a result, the linguistic characteristics of LLM-generated simplifications remain largely unexplored.

We investigate two adaptation methods for LLM-based text simplification: fine-tuning and prompting. To investigate these gaps, we ask three research questions:

RQ1: How do adaptation methods perform across simplification domains?

RQ2: How effectively do adaptation methods support controllable simplification?

RQ3: How do LLM-generated simplifications differ from human-written references?

To answer these questions, we evaluate 12 open-weight LLMs from three model families (Qwen, DeepSeek-R1-Distill-Qwen, and Llama; 1B–32B parameters) on four English datasets spanning Wikipedia (ASSET; Alva-Manchego et al. 2020),

medical (Med-Easi; Basu et al. 2023), scientific (CLEF 2025; Bakker et al. 2025), and news (Newsela; Xu et al. 2015) domains. We compare two practical approaches to text simplification: QLoRA fine-tuning of base models and optimized prompting of instruction-tuned models. We additionally report a controlled four-way analysis to examine the effects of checkpoint type and adaptation method (Section 4.1; Appendix Table 3). Finally, we perform a targeted manual error analysis to validate our findings. SIMPLE extends prior benchmarks (Kew et al., 2023; Qiang et al., 2025; Scalercio et al., 2025) through comparisons across domains and model scales.¹

2 Related Work

2.1 Prompting and Fine-Tuning

Prompting-based approaches have shown strong performance on text simplification tasks (Feng et al., 2023; Kew et al., 2023; Qiang et al., 2025; Zhang et al., 2026). Fine-tuning methods, including SFT, RLHF, and parameter-efficient techniques such as LoRA, have also been applied to medical and lexical simplification (Rahman et al., 2024; Baez and Saggion, 2023). While recent studies suggest that fine-tuning can outperform prompting (Cohen et al., 2026), they also report issues such as overfitting, reduced diversity, and poor cross-domain generalization (Devaraj et al., 2022). Systematic comparisons across domains and model scales remain limited.

2.2 Controllable Text Simplification

Controllable text simplification aims to generate outputs at specified readability levels. Existing LLM-based TS studies have primarily focused on overall simplification quality, with limited attention to controllability (Kew et al., 2023; Qiang et al., 2025; Cohen et al., 2026). Although datasets such as Newsela provide aligned simplifications across multiple reading levels (Xu et al., 2015), most prior work, including BLESS (Kew et al., 2023), treats them as standard benchmarks rather than for evaluating controllable simplification. Consequently, the effectiveness of different adaptation strategies across target readability levels remains unclear.

2.3 Evaluation

Text simplification is commonly evaluated using automatic metrics and human judgments. Widely used metrics such as SARI and FKGL correlate weakly with human judgments and offer limited discriminability for high-quality outputs (Alva-Manchego et al., 2021; Kew et al., 2023; Wu and Arase, 2026). LENS shows stronger alignment with human evaluation but remains a holistic measure with limited interpretability at the level of specific simplification operations (Maddela et al., 2023). Human evaluation assesses fluency, meaning preservation, and simplicity, but is costly, difficult to scale, and exhibits low inter-annotator agreement (Alva-Manchego et al., 2021; Wu and Arase, 2026). Consequently, recent work has explored fine-grained analyses of simplification outputs, including edit-level annotation frameworks such as SALSA (Heineman et al., 2023) and linguistic analyses based on coherence, lexical sophistication, and discourse and syntactic features (Huynh and McNamara, 2025; Reinhart et al., 2025). These studies indicate that linguistic analysis captures distinctions between LLM-generated and human-written simplifications not reflected in standard metrics.

3 Methodology

Experiments are conducted on NVIDIA RTX 4090 (24GB) or RTX 3090 (24GB) GPUs. We fix random seeds, enable deterministic CUDA settings, and keep software versions consistent across runs; a hardware-consistency check is reported in Appendix D. We release the code, configuration files, and evaluation scripts with this paper.

3.1 Datasets

We evaluate four datasets covering Wikipedia, medical, scientific, and news domains (Table 1).

ASSET (Alva-Manchego et al., 2020) is a Wikipedia-based benchmark covering lexical substitution, sentence splitting, and deletion. We use the 2,000-sentence pool for training and the official 359-sentence test set (10 references per source).

Med-Easi (Basu et al., 2023) focuses on medical simplification into lay language, involving substitution, deletion, insertion, and elaboration. We follow the official split.

CLEF (Bakker et al., 2025) contains biomedical texts paired with simplified summaries. We use the official Task 1 training set for training and 932 sentence pairs for testing.

¹We make our code and experimental configurations available at <https://github.com/mengzixian/SIMPLE>.

Dataset	Domain	Train	Test	#R	TER
ASSET	Wikipedia	2,000	359	10	16.79
Med-Easi	Medical	1,397	300	1	81.82
CLEF	Science	14,719	932	1	81.24
		11,627	4,984	1	55.84
Newsela	News	14,085	6,037	1	86.03
		13,923	5,968	1	116.81
		11,260	4,826	1	151.57

Table 1: Dataset statistics. #R denotes the number of references per source sentence, and TER denotes the average edit distance between source and reference texts.

Newsela (Xu et al., 2015) consists of professionally rewritten news articles across four reading levels for controllable simplification analysis. We use a 60%/40% train/test split at each readability level.

3.2 LLMs

We evaluate 12 open-weight LLMs from three model families—Qwen2.5, DeepSeek-R1-Distill-Qwen (abbreviated DS-R1 in tables), and Llama—ranging from 1B to 32B parameters. Model details and training pipelines are summarized in Appendix Table 10.

Qwen2.5 (1.5B, 7B, 14B, and 32B) is a general-purpose multilingual LLM family trained on up to 18 trillion tokens and has demonstrated strong performance across a wide range of NLP tasks. DeepSeek-R1-Distill-Qwen (1.5B, 7B, 14B, and 32B) consists of reasoning-oriented models distilled from DeepSeek-R1, incorporating chain-of-thought reasoning capabilities through supervised distillation. Llama includes Llama-3.2-1B, Llama-3.2-3B, Llama-3.1-8B, and Llama-2-13B, representing widely adopted open-weight foundation models trained on large-scale web corpora.

To ensure a fair and practical comparison between adaptation strategies, we adopt their standard usage settings. Fine-tuning is performed on base models to isolate performance gains from parameter updates, without the influence of prior instruction tuning. In contrast, prompting is applied to instruction-tuned models, which are optimized for instruction following and represent the standard and most effective setup for prompt-based inference. This design reflects real-world deployment practices while ensuring that each method is evaluated under its appropriate configuration. Because checkpoint type is confounded with adaptation method in this comparison, we also conduct a controlled four-way experiment using Qwen2.5-7B

on ASSET (Appendix Table 3).

3.3 Evaluation

We evaluate model outputs using three components: automatic metrics, linguistic analysis, and manual evaluation. Automatic evaluation includes SARI (Xu et al., 2016), FKGL (Kincaid et al., 1975), BERTScore (Zhang et al., 2019), and LENS (Maddala et al., 2023), a learned neural metric trained on human judgments of simplification quality. Following Kew et al. (2023), leave-one-out scoring is used for multi-reference datasets, while single-reference datasets (Med-Easi, CLEF, and Newsela) are evaluated against their provided references. We further analyze outputs using Coh-Metrix (Graesser et al., 2004), extracting features spanning syntactic, lexical, discourse, and readability dimensions. The manual evaluation examines specific errors in content preservation and simplification adequacy.

Specifically, we use Qwen2.5-7B as a representative model and sample 70 cases: 10 from each of ASSET, CLEF, and Med-Easi and 10 from each Newsela readability level. Whereas the automatic metrics measure reference-based edits, readability, semantic similarity, or overall quality, the manual evaluation directly examines two dimensions adapted from Vendeville et al. (2025): *content preservation*, including omission, distortion, and unsupported information, and *simplification*, including over- and incomplete simplification. Three annotators evaluate fine-tuning outputs, prompting outputs, and human references. For each output type, Fleiss’ κ is computed over 350 case–category items, each rated by all three annotators. Agreement between output types is measured using Cohen’s κ and averaged over all nine cross-annotator pairings.

3.4 Prompting and Fine-Tuning

We use MIPROv2 (Opsahl-Ong et al., 2024) within the DSPy framework (Khattab et al., 2023), as preliminary experiments showed consistently stronger performance than BootstrapFewShot across all datasets. MIPROv2 jointly optimizes task instructions and few-shot demonstrations through iterative search guided by a task-specific objective. For prompt optimization only, we randomly sample 800 instances from each dataset’s training set and split them into 600 optimization and 200 validation examples; this subset is distinct from the full training data used for fine-tuning (Table 1). SARI is used as the optimization objective, and candidate

prompts producing outputs with a SARI score of zero are excluded during optimization. We use a single bootstrapped demonstration during prompt search. The resulting optimized prompts are then used for inference with chain-of-thought prompting.

We fine-tune all models using QLoRA (Dettmers et al., 2023) with 4-bit quantization and BF16 mixed-precision training on the full training split of each dataset. Optimization is performed using AdamW (learning rate = 5×10^{-5}), LoRA rank 8, LoRA alpha 16, batch size 2, gradient accumulation of 4, a maximum input length of 256 tokens, and 2 training epochs. Pilot runs indicated that training and validation loss plateau after roughly 400–500 optimization steps, with negligible gains from a third epoch; we therefore use 2 epochs for a controlled comparison rather than per-model hyperparameter search. The same hyperparameter configuration is applied across all models and datasets. During inference, LoRA adapters are merged into the base model and served via vLLM with a maximum generation length of 150 tokens. The instructions used for both methods and a representative MIPROv2-optimized prompt are provided in Appendix A.

3.5 Post-processing

Sentence simplification requires outputs to be well-formed sentences. To ensure fair and consistent comparison between prompting and QLoRA fine-tuning, we apply a unified post-processing pipeline to all outputs. We remove role markers, special tokens, reasoning traces (e.g., <think> blocks), and prompt template artifacts, and normalize whitespace and quotation marks. We then apply sentence-level filtering to retain the first complete sentence satisfying basic grammatical constraints, reducing noise from partial or malformed generations.

This first-sentence filter may discard legitimate sentence splitting, especially on ASSET. On a controlled ASSET diagnostic with Qwen2.5-14B fine-tuning, first-sentence outputs score higher than full cleaned outputs (SARI 44.07 vs. 42.01; BERTScore 0.93 vs. 0.88; LENS 64.68 vs. 13.93; Appendix Table 4). Manual inspection shows that most removed content is decoding artifacts rather than valid continuations, for example “*His next work, Saturday, takes place on an eventful day in the life of a successful neurosurgeon. user You are a text simplification assistant...*”. The first-sentence filter keeps only the initial sentence and

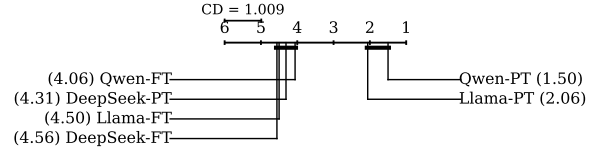


Figure 1: Critical difference (CD) diagram comparing six conditions defined by three model families (LLaMA, Qwen, DeepSeek) under two adaptation methods on LENS (Nemenyi post-hoc test, $CD = 1.009$, $N = 16$, $\alpha = 0.05$). Connected groups are not significantly different.

drops the leaked prompt. Thus the filter mainly reduces evaluation noise, but it is not an ideal general solution and may underestimate splitting-based simplification. Future work could reduce reliance on this heuristic by using inference-level stop tokens, stricter generation templates, and length-constrained decoding.

3.6 Baselines

Following Kew et al. (2023), we use MUSS (Martin et al., 2022) in both supervised (muss_en_wikilarge_mined) and unsupervised (muss_en_mined) settings. MUSS is a BART-based simplification model trained on automatically mined paraphrase corpora and controlled through simplification-specific tokens. We use the released checkpoints with default inference settings.

4 Results

4.1 Automatic Evaluation Results

We report results on ASSET in Table 2 and Figure 2; full results for Med-Easi, CLEF, and Newsela are in Appendix C (Table 6).

Fine-tuning performs better on Wikipedia-style simplification, while prompting is more effective in non-Wikipedia settings. On ASSET, fine-tuned models achieve slightly higher average SARI than prompting, with mixed LENS results across families. In contrast, prompting yields higher LENS scores on Med-Easi, CLEF, and Newsela across model families, with more pronounced gains for Qwen and Llama. Although fine-tuning often matches or exceeds SARI, it does not surpass the gold-reference LENS on non-Wikipedia datasets. Prompting achieves higher LENS than the gold-reference baseline on CLEF and Newsela (67.12 and 68.76, respectively) and approaches it on Med-Easi (65.41), while DeepSeek shows smaller differences between the two pipelines. Relative to

ASSET		Fine-tuning				Prompting			
		SARI↑	FKGL↓	BERT↑	LENS↑	SARI↑	FKGL↓	BERT↑	LENS↑
Baselines	Gold references	43.86	6.58	0.94	61.29	43.86	6.58	0.94	61.29
	MUSS_mined	41.81	8.18	0.94	57.34	41.81	8.18	0.94	57.34
	MUSS_wiki_mined	44.25	5.29	0.94	66.03	44.25	5.29	0.94	66.03
LLMs	Qwen2.5-1.5B	43.34	6.74	0.93	63.52	42.71	7.83	0.93	63.47
	Qwen2.5-7B	43.83	6.76	0.93	64.20	44.70	6.76	0.94	66.81
	Qwen2.5-14B	44.07	6.67	0.93	64.68	44.80	7.29	0.94	67.48
	Qwen2.5-32B	44.49	6.80	0.93	64.88	43.48	7.34	0.94	65.52
	DS-R1-1.5B	42.42	7.72	0.93	57.08	38.29	7.75	0.92	47.73
	DS-R1-7B	43.13	6.56	0.93	61.97	40.83	8.39	0.94	59.06
	DS-R1-14B	43.87	6.22	0.93	65.41	44.32	8.06	0.94	62.15
	DS-R1-32B	43.94	6.36	0.93	65.00	43.46	7.14	0.93	61.40
	Llama-3.2-1B	42.83	6.75	0.93	62.03	36.53	5.14	0.91	57.09
	Llama-3.2-3B	43.78	6.32	0.93	64.56	36.14	5.26	0.91	63.49
	Llama-3.1-8B	43.98	6.53	0.93	64.44	39.68	5.15	0.92	68.70
	Llama-2-13B	44.45	6.64	0.93	63.85	41.62	7.27	0.93	64.24

Table 2: Automatic evaluation results on ASSET under fine-tuning and prompting. Bold denotes the best result within each adaptation method.

gold references on Med-Easi and CLEF, fine-tuned Qwen and DeepSeek models consistently obtain lower FKGL, whereas Llama prompting often yields lower FKGL than Llama fine-tuning (Table 2 and Table 6); we therefore treat this readability pattern as model-family-specific rather than universal.

Because the main comparison confounds check-point type with adaptation method, we conduct a controlled four-way analysis on Qwen2.5-7B and ASSET (Appendix Table 3). Base+FT substantially outperforms Base+Prompt (SARI 43.83 vs. 36.60; LENS 64.20 vs. 54.95), while Instruct+Prompt achieves the best LENS (66.81) and slightly outperforms Instruct+FT. Thus, both initialization and adaptation strategy matter, and our main results should be read as a comparison of deployment pipelines rather than as evidence that one strategy is universally superior.

Adaptation pipeline has a larger impact than model scale, with variations across model families. On ASSET, SARI increases with model size under fine-tuning for Qwen and DeepSeek, while Llama shows limited scaling effects. Under prompting, Qwen peaks at 14B before declining at 32B, whereas DeepSeek prompting consistently underperforms fine-tuning across scales (e.g., 38.29 vs. 42.42 SARI for 1.5B). The Nemenyi post-hoc test (CD = 1.009; Figure 1) further shows that Qwen- and Llama-prompting variants rank higher in LENS than most fine-tuning settings, while DeepSeek-prompting is not significantly different from fine-tuned variants, suggesting weaker separation between methods for this family. Overall, the choice of adaptation pipeline yields larger differences than

model scaling.

SARI and LENS capture different aspects of simplification quality. On Newsela, fine-tuning achieves higher SARI (41.78 vs. 39.26) but lower LENS across model families, indicating divergence between lexical overlap-based and learned evaluation signals. BERTScore remains stable across methods (0.89–0.94), with a modest decline on Med-Easi, consistent with increased difficulty in preserving semantic content in domain-specific simplification.

4.2 Human Evaluation

The manual analysis shows distinct error patterns (Appendix Figure 6). Fine-tuning has the highest omission (61.0%), distortion (21.0%), and oversimplification rates (41.4%), indicating aggressive compression that sometimes removes or alters important information. Prompting has substantially lower rates in these categories (31.4%, 7.1%, and 2.9%) and also the lowest unsupported-information rate (2.9%). Incomplete simplification is relatively uncommon for all three output types. This distinction is especially important in medical and scientific domains, where lower readability may result from deleting essential content (Devaraj et al., 2022).

Inter-annotator Fleiss’ κ is 0.946, 0.745, and 0.886 for fine-tuning, prompting, and human references, respectively. Across the nine cross-annotator pairings, mean Cohen’s κ is 0.238 for fine-tuning–prompting, 0.528 for fine-tuning–human references, and 0.267 for prompting–human references. These results suggest that fine-tuning more closely

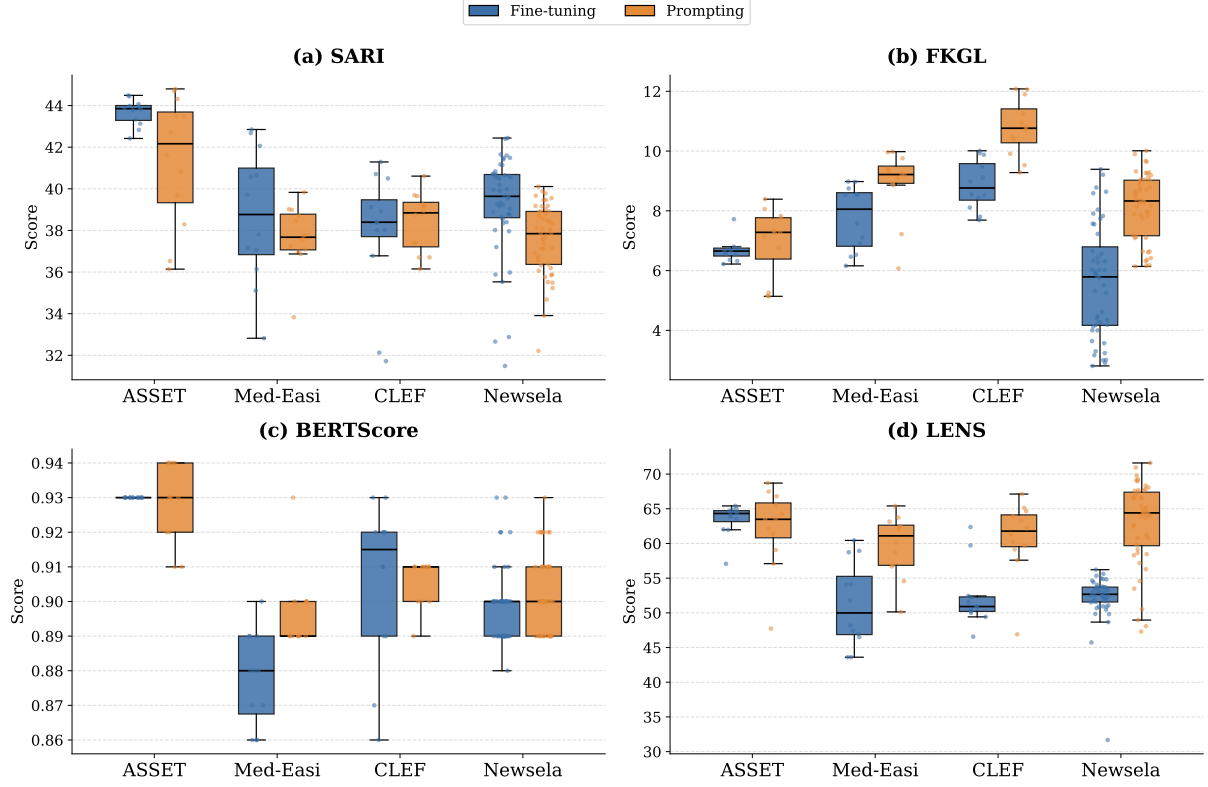


Figure 2: Distributions of SARI, FKGL, BERTScore, and LENS scores for all 12 LLMs across four datasets under fine-tuning and prompting. The central line indicates the median.

resembles the error pattern of human references, largely because both exhibit stronger information compression. However, this similarity does not indicate higher quality: prompting produces substantially fewer errors and is more reliable under our criteria. The manual analysis broadly aligns with the automatic results, as both indicate an advantage for prompting on specialized-domain datasets, despite assessing different aspects of simplification quality.

4.3 Controllable Simplification

We evaluate controllability across the four Newsela readability levels, as indicated in Figure 3 and Appendix Table 7.

Fine-tuning better captures relative readability variation across levels, while prompting shows limited controllability. Fine-tuned models exhibit a monotonic decrease in FKGL from L1 to L4, closely following the trend of the gold references, indicating sensitivity to target readability levels and the ability to model relative difficulty transitions. However, FKGL remains consistently lower than the gold references across all levels, suggesting a tendency toward stronger simplification. SARI remains relatively stable across readability levels,

showing limited sensitivity to graded difficulty signals. In contrast, prompting produces nearly flat FKGL values across levels (≈ 7 – 10), with weaker alignment to target readability progression. The gap to the gold references increases at higher difficulty levels, indicating reduced alignment under stricter readability settings. This suggests that optimization based on SARI alone provides limited signal for controlling output difficulty.

4.4 Analysis of Edit Operations

Following Vázquez-Rodríguez et al. (2021), we compute token-level KEEP, INSERT, DELETE, and REPLACE proportions using a Wagner-Fischer alignment (diffib), visualized with EASSE (Alva-Manchego et al., 2019). Figure 4 presents results on ASSET; Appendix Figures 7–12 report Med-Easi, CLEF, and Newsela.

Gold references and MUSS exhibit consistent editing patterns, while LLM-based methods show method-specific biases. Across datasets, gold references and MUSS variants show high KEEP and REPLACE rates with near-zero INSERT. On Newsela, gold references exhibit increasing REPLACE and DELETE with decreasing readability levels, accompanied by reduced KEEP, reflecting more ag-

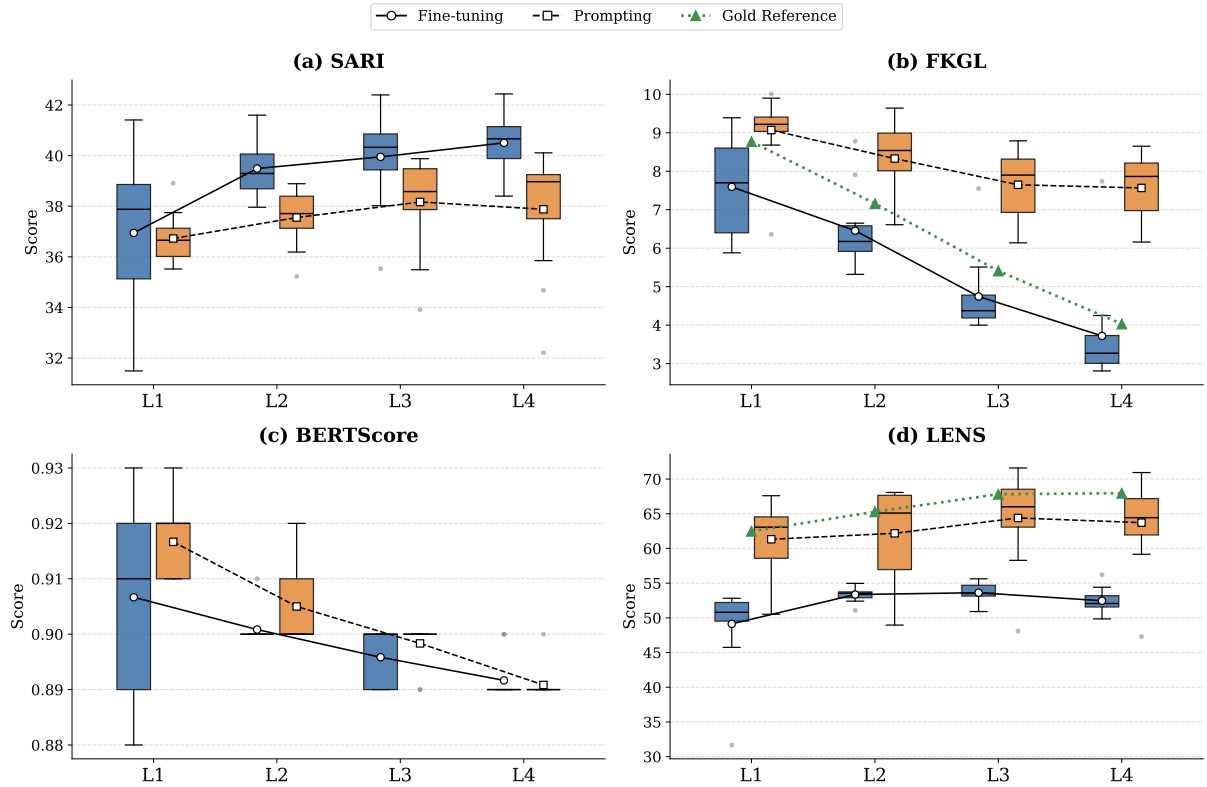


Figure 3: Distributions of SARI, FKGL, BERTScore, and LENS scores across the four Newsela readability levels (L1-L4) under fine-tuning and prompting.

gressive editing at lower difficulty levels.

Fine-tuning and prompting exhibit distinct but non-adaptive edit behaviors. Fine-tuned models consistently produce higher REPLACE rates and slightly lower KEEP rates than gold references, with stronger effects on CLEF and Med-Easi, consistent with the over-simplification trends observed in Section 4.2. In contrast, prompting-based models generally preserve KEEP rates closer to gold references, but smaller models (Llama-3.2-1B/3B) exhibit elevated DELETE rates across multiple datasets, aligning with their lower LENS scores. Neither adaptation method shows systematic changes in edit distributions across Newsela readability levels.

Statistical comparisons confirm systematic differences between methods. We conduct paired t-tests with Bonferroni correction across four edit operations ($\alpha = 0.0125$). On CLEF, prompting yields significantly higher DELETE proportions than fine-tuning (FT = 0.025, PT = 0.101, $p < 0.001$). On Newsela, differences become more pronounced across levels: at L1, only INSERT differs significantly, while at L4, KEEP, INSERT, and DELETE are all significant. At L4, fine-tuning pro-

duces higher DELETE rates (FT = 0.255 vs. PT = 0.163, $p < 0.001$), whereas prompting maintains higher KEEP rates (FT = 0.239 vs. PT = 0.355, $p < 0.001$), indicating diverging editing behaviors under stricter readability constraints.

These patterns align with FKGL trends in Section 4.3, where fine-tuning shows stronger simplification pressure. Importantly, these edit-level differences do not contradict the higher LENS scores of prompting in specialized domains (Section 4.1), as LENS captures overall quality while edit distributions reflect localized transformation behavior. Representative outputs illustrating these contrasts are shown in Appendix B.

4.5 Linguistic Characteristics of Generated Simplifications

We use Coh-Metrix (Graesser et al., 2004) to analyze generated simplifications along four dimensions: syntactic structure, lexical selection, logical coherence, and overall reading experience. Feature values are normalized against the human reference (Rate = 1). For each model family, we select the highest-performing model according to LENS and compare its outputs with the source texts, MUSS_wiki_mined, and gold references.

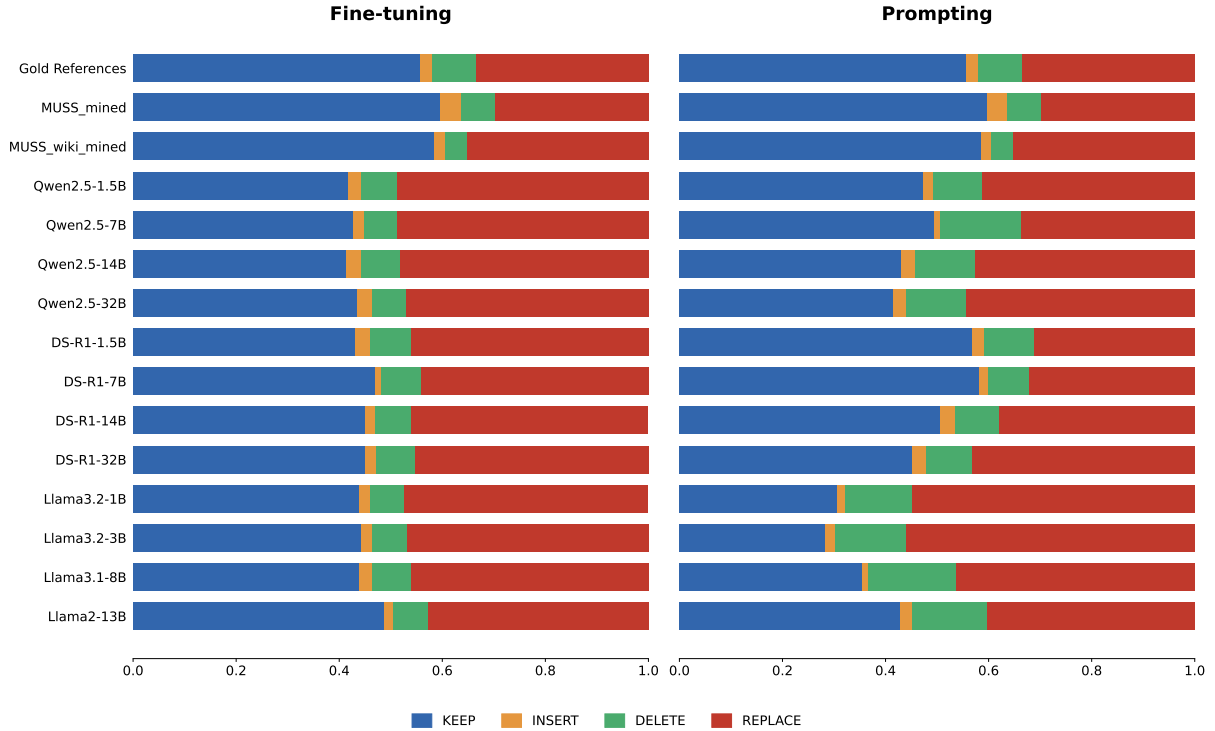


Figure 4: Distributions of Edit Operations on ASSET.

Figure 5 presents results on ASSET, while Appendix Figures 13 and 14 report Med-Easi and CLEF.

Syntactic Structure. On ASSET, prompting yields higher syntactic similarity values than the gold references, while both adaptation methods are generally closer to gold references on Med-Easi and CLEF. Parse-tree depth and phrasal density show relatively limited variation across methods and datasets.

Lexical Selection. Word frequency, age of acquisition, and concreteness remain close to gold-reference values across datasets, with only minor differences between adaptation methods. Lexical diversity is consistently higher under prompting, with the largest divergence observed on Med-Easi.

Logical Coherence. Content-word overlap exceeds gold-reference values on ASSET for both methods, particularly under prompting, but falls below gold-reference values on Med-Easi and CLEF. A similar pattern is observed for causal cohesion, suggesting substantial variation in coherence-related features across domains.

Overall Reading Experience. Sentence length, word length, and FKGL generally remain close to gold-reference values across datasets, although prompting exhibits greater variability across model families. Narrativity is higher than gold-reference

values on ASSET but closer to, or below, gold-reference values on Med-Easi and CLEF.

Overall, linguistic differences between adaptation methods are smaller than differences across datasets. Because datasets also differ substantially in rewrite difficulty (TER; Table 1), these cross-dataset gaps likely reflect both domain characteristics and task difficulty rather than domain alone. Nevertheless, fine-tuned outputs generally remain closer to gold-reference profiles, while prompting exhibits greater variation across both model families and datasets.

5 Discussion

RQ1: How do adaptation methods perform across simplification domains? Our results indicate that pipeline effectiveness depends on both dataset characteristics and rewrite difficulty. Under the practical Base+FT vs. Instruct+Prompt comparison, fine-tuning performs more strongly on the lower-TER Wikipedia ASSET setting, whereas prompting generally achieves better quality–readability trade-offs on higher-TER medical, scientific, and news data. The controlled four-way analysis (Appendix Table 3) further shows that these gaps reflect the joint effect of checkpoint initialization and adaptation method, rather than adaptation strategy alone. These findings reinforce

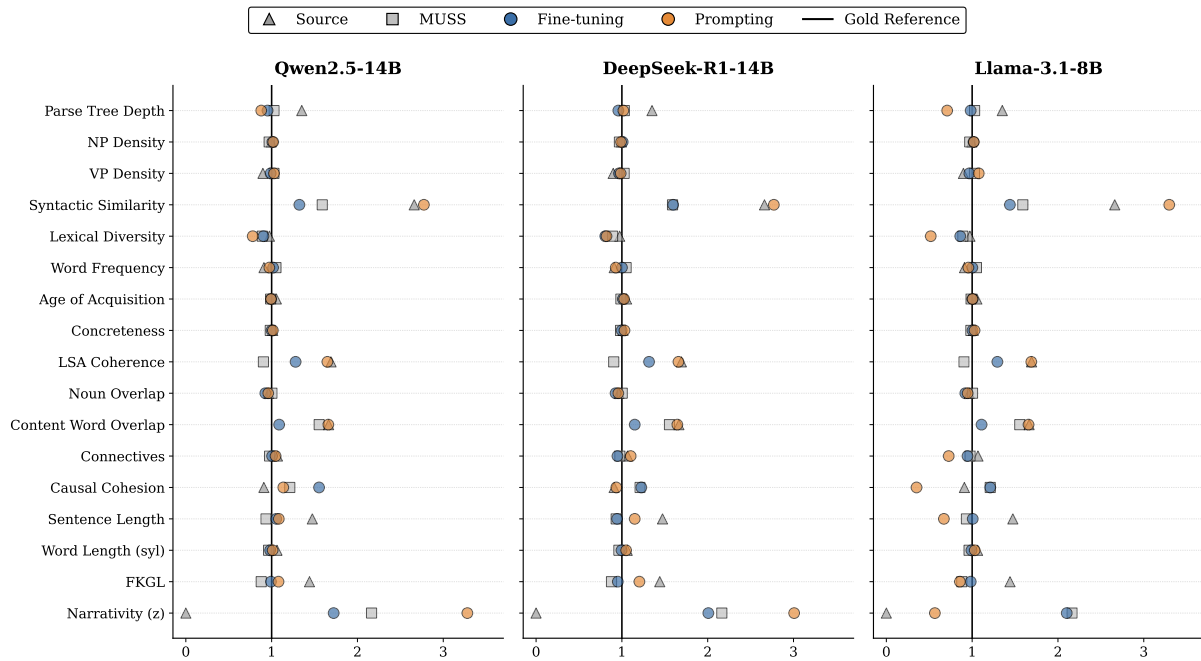


Figure 5: Linguistic feature profiles of the best-performing model from each LLM family on ASSET. See Appendix Table 9 for variable definitions.

previous observations that different evaluation criteria may favor different systems and that simplification quality cannot be fully characterized by a single metric (Asthana et al., 2024; Qiang et al., 2025). Across datasets, the choice of adaptation pipeline has a larger impact than model scale, suggesting that selecting an appropriate pipeline may be more important than increasing parameter count under fixed computational budgets.

RQ2: How effectively do adaptation methods support controllable simplification? The two adaptation methods exhibit distinct controllability characteristics. Fine-tuned models better track the relative readability progression represented by the Newsela levels, indicating stronger alignment with graded difficulty changes. However, their outputs remain consistently simpler than the corresponding human references. In contrast, prompting shows comparatively limited variation across target readability levels, resulting in weaker alignment with the intended difficulty progression. Consistent with recent work on readability-controlled generation (Barayan et al., 2025), these results suggest that controllable simplification remains challenging for current LLM adaptation methods and may require more explicit readability signals during adaptation.

RQ3: How do LLM-generated simplifications differ from human-written references? Model

outputs differ most from human references in syntactic and discourse features, while lexical characteristics are more similar. Fine-tuning is generally closer to human references, whereas prompting varies more across datasets and model families, showing that these differences are method- and dataset-dependent (Huynh and McNamara, 2025; Reinhart et al., 2025).

6 Conclusion

We compare fine-tuning and prompting across domains and model scales. Performance depends more on domain, rewrite difficulty, checkpoint type, and adaptation method than scale. Linguistic and manual analyses find fewer prompting errors but greater similarity between fine-tuning and human syntactic, discourse, and information-compression patterns. Future work should develop methods targeting specific readability levels while preserving essential content, with validation by target readers.

Limitations

This study has several limitations. First, although we evaluate four datasets spanning Wikipedia, medical, scientific, and news domains, simplification requirements vary substantially across text types. Additional domains, such as legal, educational, and governmental texts, would help assess the generalizability of our findings. Moreover, our ex-

periments focus on sentence-level simplification and do not examine document-level simplification phenomena, including discourse organization and cross-sentence coherence, which may further affect the behavior of different adaptation methods. Second, our experiments focus exclusively on open-weight models and exclude large proprietary systems such as OpenAI GPT-4-class models and Anthropic Claude-class models. In addition, the largest model considered is 32B parameters; the behavior of larger models remains an open question. Third, controllability is assessed using readability-related metrics rather than direct human evaluation of target difficulty levels. Our prompt optimization also uses SARI as its sole objective, although SARI may not adequately capture target readability or content preservation. Future work should explore multi-objective optimization that jointly considers simplification quality, readability control, and meaning preservation. Our manual error analysis is also limited to 70 Qwen2.5-7B cases, including only 10 cases from each of ASSET, Med-Easi, and CLEF; dataset-specific differences in these small subsets should therefore be treated as descriptive trends. Future work could combine larger-scale error annotation with direct judgments from target readers. Finally, our controlled factorial analysis is limited to Qwen2.5-7B on ASSET. Extending all four checkpoint-adaptation combinations across model families, scales, and domains remains an important direction for future work.

Acknowledgments

We thank the anonymous reviewers for their constructive feedback. We also thank the annotators for their careful work on the manual evaluation.

Ethics Statement

This work follows the ACL Ethics Policy and aims to improve the transparency and reliability of evaluation for text simplification. The models, datasets, and computational resources are used in accordance with their applicable access conditions and licenses. The manual evaluation concerns system outputs and does not collect sensitive personal information. We recognize that simplified texts, particularly in medical and scientific domains, may omit or distort important information; the generated outputs are intended for research evaluation rather than clinical or professional decision-making.

References

- Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. 2020. [Asset: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4668–4679.
- Fernando Alva-Manchego, Louis Martin, Carolina Scarton, and Lucia Specia. 2019. [Easse: Easier automatic sentence simplification evaluation](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 49–54.
- 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.
- Sumit Asthana, Hannah Rashkin, Elizabeth Clark, Fantine Huot, and Mirella Lapata. 2024. [Evaluating llms for targeted concept simplification for domain-specific texts](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6208–6226.
- Anthony Baez and Horacio Saggion. 2023. [Lslama: Fine-tuned llama for lexical simplification](#). In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 102–108.
- Jan Bakker, Benjamin Vendeville, Liana Ermakova, and Jaap Kamps. 2025. [Overview of the clef 2025 simple-text task 1: Simplify scientific text](#). *Working Notes of CLEF*, pages 3147–3162.
- Abdullah Barayan, Jose Camacho-Collados, and Fernando Alva-Manchego. 2025. [Analysing zero-shot readability-controlled sentence simplification](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6762–6781.
- Paul-Gerhard Barbu, Adrianna Lipska-Dieck, and Lena Lindner. 2025. [Easyjon at tsar 2025 shared task evaluation of automated text simplification with llm-as-a-judge](#). In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility and Readability (TSAR 2025)*, pages 173–182.
- Chandrayee Basu, Rosni Vasu, Michihiro Yasunaga, and Qian Yang. 2023. [Med-easi: Finely annotated dataset and models for controllable simplification of medical texts](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14093–14101.
- Eilam Cohen, Itamar Bul, Danielle Inbar, and Omri Loewenbach. 2026. [Simplify-this: A comparative analysis of prompt-based and fine-tuned llms](#). *arXiv preprint arXiv:2601.05794*.

- Scott A Crossley. 2025. [Developing linguistic constructs of text readability using natural language processing](#). *Scientific Studies of Reading*, 29(2):138–160.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. [Qlora: Efficient finetuning of quantized llms](#). *Advances in neural information processing systems*, 36:10088–10115.
- Ashwin Devaraj, William Sheffield, Byron C Wallace, and Junyi Jessie Li. 2022. [Evaluating factuality in text simplification](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7331–7345.
- Yutao Feng, Jipeng Qiang, Yun Li, Yunhao Yuan, and Yi Zhu. 2023. [Sentence simplification via large language models](#). *arXiv preprint arXiv:2302.11957*.
- Arthur C Graesser, Danielle S McNamara, Max M Louwerse, and Zhiqiang Cai. 2004. [Coh-metrix: Analysis of text on cohesion and language](#). *Behavior research methods, instruments, & computers*, 36(2):193–202.
- David Heineman, Yao Dou, Mounica Maddela, and Wei Xu. 2023. [Dancing between success and failure: Edit-level simplification evaluation using salsa](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3466–3495.
- Linh Huynh and Danielle S McNamara. 2025. [Natural language processing as a scalable method for evaluating educational text personalization by llms](#). *Applied Sciences*, 15(22):12128.
- Bahar Ilgen and Georges Hattab. 2025. [Toward human-centered readability evaluation](#). In *Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP)*, pages 263–273.
- Itay Itzhak, Gabriel Stanovsky, Nir Rosenfeld, and Yonatan Belinkov. 2024. [Instructed to bias: Instruction-tuned language models exhibit emergent cognitive bias](#). *Transactions of the Association for Computational Linguistics*, 12:771–785.
- 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.
- Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T Joshi, and Hanna Moazam. 2023. [Dspy: Compiling declarative language model calls into self-improving pipelines](#). *arXiv preprint arXiv:2310.03714*.
- 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](#). Report, Naval Air Station Memphis.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. [Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.
- 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.
- Louis Martin, Angela Fan, Éric Villemonte 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.
- Krista Opsahl-Ong, Michael J Ryan, Josh Purtell, David Broman, Christopher Potts, Matei Zaharia, and Omar Khattab. 2024. [Optimizing instructions and demonstrations for multi-stage language model programs](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9340–9366.
- Jipeng Qiang, Minjiang Huang, Yi Zhu, Yunhao Yuan, Chaowei Zhang, and Kui Yu. 2025. [Redefining simplicity: Benchmarking large language models from lexical to document simplification](#). *arXiv preprint arXiv:2502.08281*.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *Journal of Machine Learning Research*, 21(140):1–67.
- Md Mushfiqur Rahman, Mohammad Sabik Irbaz, Kai North, Michelle S Williams, Marcos Zampieri, and Kevin Lybarger. 2024. [Health text simplification: An annotated corpus for digestive cancer education and novel strategies for reinforcement learning](#). *Journal of Biomedical Informatics*, 158:104727.
- Alex Reinhardt, Ben Markey, Michael Laudenbach, Kachad Pantusen, Ronald Yurko, Gordon Weinberg, and David West Brown. 2025. [Do llms write like humans? variation in grammatical and rhetorical styles](#). *Proceedings of the National Academy of Sciences*, 122(8):e2422455122.

Arthur Mariano Rocha De Azevedo Scalercio, Elvis A De Souza, Maria José B Finatto, and Aline Paes. 2025. [Evaluating llms for portuguese sentence simplification with linguistic insights](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24452–24477.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. [Bleu is not suitable for the evaluation of text simplification](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 738–744.

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.

Benjamin Vendeville, Liana Ermakova, and Pierre De Loor. 2025. [Resource for error analysis in text simplification: New taxonomy and test collection](#). In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3723–3732. Association for Computing Machinery.

Xuanxin Wu and Yuki Arase. 2026. [An in-depth evaluation of large language models in sentence simplification with error-based human assessment](#). *ACM Transactions on Intelligent Systems and Technology*, 17(4):1–25.

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

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. [Optimizing statistical machine translation for text simplification](#). *Transactions of the Association for Computational Linguistics*, 4:401–415.

Jingshen Zhang, Xin Ying Qiu, Lifang Lu, Zhuhua Huang, Yutao Hu, Yuechang Wu, and JunYu Lu. 2026. [Let’s simplify step by step: Guiding llm towards multilingual unsupervised proficiency-controlled sentence simplification](#). In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 5274–5290.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. [Bertscore: Evaluating text generation with bert](#). *arXiv preprint arXiv:1904.09675*.

Appendix

A Prompting and Fine-Tuning Procedures

A.1 Prompting

For prompting-based methods, we use task-specific instructions and prompts optimized for each dataset.

A typical prompt is: “*Rewrite the following complex sentence into a simpler version.*”

We employ MIPROv2 (Opsahl-Ong et al., 2024) within the DSPy framework (Khattab et al., 2023), which automatically optimizes prompts through iterative search. The optimization process consists of four stages:

1. **Example sampling:** Six representative few-shot examples are selected from the training set.
2. **Instruction generation:** Candidate instructions are generated based on the sampled examples.
3. **Bayesian optimization:** Instruction–example combinations are iteratively optimized over 13 rounds using SARI as the objective.
4. **Final prompt selection:** The highest-performing prompt configuration is selected for inference.

A representative instruction generated for Llama-2-13B on CLEF is shown below:

“*Use the information below to learn about a task and generate an instruction that improves simplification performance.*”

The selected instruction does not explicitly specify the simplification task, but instead repeats a high-level prompt used for instruction generation. This indicates that the model does not always generate fully task-specific instructions during candidate generation. Despite this, the bootstrap demonstration remains effective and is associated with improved simplification performance.

A representative demonstration is shown below:

Source: *We included one randomized controlled trial (RCT) on pharmacological interventions; the trial included 95 participants.*

Reasoning: *The included RCT provided limited evidence on pharmacological interventions for bronchiolitis.*

Simplified: *One RCT studied pharmacological interventions for bronchiolitis, including 95 participants.*

A.2 Fine-tuning

We use task-specific instructions during training and inference. The training instruction is:

“You are a text simplification assistant. Rewrite the following sentence into a simpler and clearer version, while preserving the original meaning.”

Models are trained on parallel (source, reference) pairs using QLoRA (Dettmers et al., 2023) with the hyperparameters described in Section 3.4. At inference time, the fine-tuned adapter is merged into the base model and evaluated using the same task instruction.

B Examples

We present representative outputs from Qwen2.5-14B on ASSET and across the four Newsela readability levels.

B.1 ASSET

Source: *One side of the armed conflicts is composed mainly of the Sudanese military and the Janjaweed, a Sudanese militia group recruited mostly from the Afro-Arab Abbala tribes of the northern Rizeigat region in Sudan.*

Reference: *On one side of the conflicts are the Sudanese military and the Janjaweed, a Sudanese militia group. They are mostly recruited from the Afro-Arab Abbala tribes.*

Fine-tuning: *The Sudanese military and the Janjaweed (a Sudanese militia) make up one side of the armed conflicts.*

Prompting: *One side in the conflict includes the Sudanese military and a militia called Janjaweed, which is made up mostly of people from northern Sudan’s Rizeigat region.*

Fine-tuning reduces background information and produces a shorter sentence, whereas prompting retains more source details and focuses more on rephrasing.

B.2 Newsela

Level 1

Source: *While all students receive the books, they are in widest use in poor areas, where they are often the only option for teachers.*

Reference: *In those areas, the books are often the only option for teachers.*

Fine-tuning: *But they are used most widely in poor areas, where they are often the only option for teachers.*

Prompting: *The books are used most in poor areas, where they often provide the only teaching materials available.*

Level 2

Source: *The thefts emerged three or four years ago and are now “the latest, greatest thing” for organized groups seeking to steal freight, said J.J. Coughlin, vice president for law enforcement services at LoJack SCI, a supply chain protection company.*

Reference: *The thefts emerged three or four years ago. Now they are “the latest, greatest thing” for groups looking to steal goods, said J.J. Coughlin, who works at LoJack SCI, a supply chain protection company.*

Fine-tuning: *They are now “the latest, greatest thing” for criminal gangs trying to steal cargo, said J.*

Prompting: *The thefts started a few years ago and are now popular among organized groups seeking to steal freight, according to J.J. Coughlin from LoJack SCI.*

Fine-tuning produces a truncated output (“said J.”), suggesting over-editing behavior consistent with the findings in Section 4.2.

Level 3

Source: *And in 2005, a summit of world leaders adopted the doctrine of the “responsibility to protect,” which obliged the international community to move in when civilians are under attack and their governments fail to protect them.*

Reference: *And in 2005, a summit of world leaders agreed on the principle of the “responsibility to protect.”*

Fine-tuning: *In 2005, world leaders agreed on the “responsibility to protect.”*

Prompting: *In 2005, world leaders agreed to protect civilians when their governments fail to do so.*

Level 4

Source: *They included three schoolteachers from Kendallville, Ind., whose elementary school pupils sold lollipops to raise more than \$400 to contribute to the statue’s repair effort; and actor Dominic Chianese, best-known for his role on “The Sopranos” but also an accomplished tenor who sang “God Bless America” to the crowd.*

Reference: *Their pupils sold lollipops to raise money for the statue’s repairs.*

Fine-tuning: *Three teachers were there.*

Prompting: *Three schoolteachers from Kendal-*

lville raised over \$400 by selling lollipops, and actor Dominic Chianese sang “God Bless America” to help with the statue’s repair effort.

The example at Level 4 illustrates the differing behaviors of the two adaptation methods. Fine-tuning generates a substantially shorter output (“*Three teachers were there.*”), consistent with its lower FKGL scores and higher deletion rates. In contrast, prompting preserves more source content and exhibits a smaller change in readability level, consistent with the quantitative results presented in Section 4.3.

C Supplemental Results

Table 3 reports the controlled four-way comparison; Table 4 reports the post-processing diagnostic.

D Reproducibility Details

Unless otherwise noted, we use random seed 42, deterministic CUDA algorithms where available, and identical package versions for inference across GPUs. Table 5 compares RTX 4090 and RTX 3090 on Llama-3.2-1B / Med-Easi under both pipelines; differences are negligible relative to method-level gaps ($\text{SARI} \leq \pm 0.2$, $\text{LENS} \leq \pm 0.5$).

E Human Evaluation Details

Figure 6 reports the complete error-category results for the 70 evaluated cases.

Checkpoint	Adaptation	SARI↑	FKGL↓	BERT↑	LENS↑
Qwen2.5-7B-Base	Fine-tuning	43.83	6.76	0.93	64.20
Qwen2.5-7B-Base	Prompting	36.60	8.10	0.93	54.95
Qwen2.5-7B-Instruct	Fine-tuning	44.46	6.74	0.94	62.41
Qwen2.5-7B-Instruct	Prompting	44.70	6.76	0.94	66.81

Table 3: Controlled four-way comparison on ASSET (Qwen2.5-7B). Main experiments compare Base+FT vs. Instruct+Prompt.

Output setting	SARI↑	BERT↑	LENS↑
First-sentence filter	44.07	0.93	64.68
Full cleaned output	42.01	0.88	13.93

Table 4: Post-processing diagnostic on ASSET (Qwen2.5-14B fine-tuning). The large LENS drop without first-sentence filtering is caused by prompt/role leakage in generated continuations.

GPU	Adaptation	SARI↑	FKGL↓	BERT↑	LENS↑
RTX 4090	Prompting	36.37	7.32	0.89	60.89
RTX 3090	Prompting	36.47	7.60	0.89	60.50
RTX 4090	Fine-tuning	39.42	6.91	0.87	43.61
RTX 3090	Fine-tuning	39.09	7.14	0.86	43.15

Table 5: Hardware consistency check on Llama-3.2-1B / Med-Easi. The small differences between RTX 4090 and RTX 3090 indicate negligible hardware-related variation.

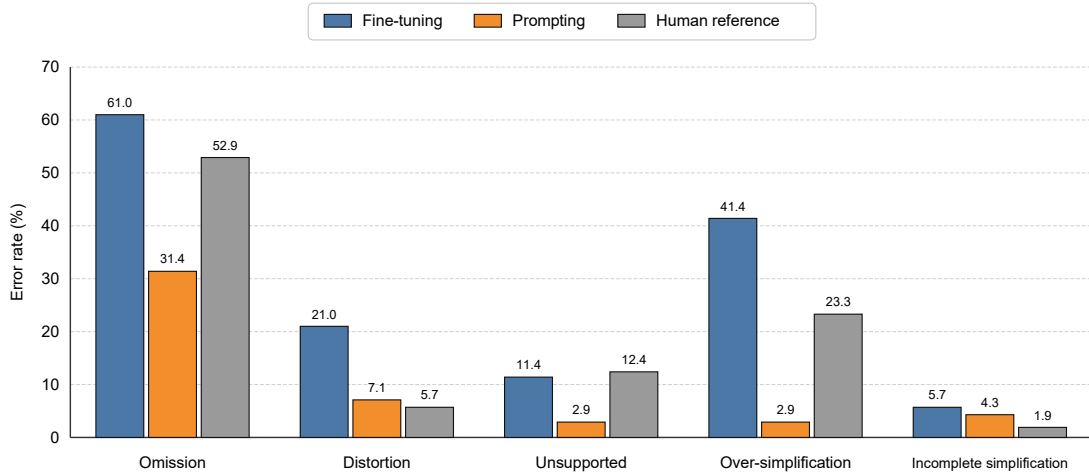


Figure 6: Manual error rates across five categories for 70 Qwen2.5-7B cases, averaged across three annotators. Categories are non-exclusive and therefore do not sum to 100%.

Med-Easi		Fine-tuning				Prompting			
		SARI↑	FKGL↓	BERT↑	LENS↑	SARI↑	FKGL↓	BERT↑	LENS↑
Baselines	Gold references	100.00	9.22	1.00	65.60	100.00	9.22	1.00	65.60
	MUSS_mined	35.02	9.04	0.90	51.20	35.02	9.04	0.90	51.20
	MUSS_wiki_mined	33.93	8.00	0.91	57.09	33.93	8.00	0.91	57.09
LLMs	Qwen2.5-1.5B	37.16	6.16	0.87	48.19	38.46	9.10	0.90	54.58
	Qwen2.5-7B	36.14	6.53	0.86	54.10	39.02	9.41	0.90	62.34
	Qwen2.5-14B	35.11	8.98	0.86	51.78	39.83	9.24	0.93	63.74
	Qwen2.5-32B	32.82	8.53	0.86	54.09	38.99	9.19	0.90	62.46
	DS-R1-1.5B	37.81	8.54	0.88	47.42	33.83	9.76	0.89	50.14
	DS-R1-7B	37.07	6.46	0.89	46.51	37.09	9.96	0.89	56.90
	DS-R1-14B	40.64	7.11	0.88	46.99	37.00	8.86	0.89	56.69
	DS-R1-32B	42.85	8.97	0.90	58.74	37.77	8.94	0.89	58.70
	Llama-3.2-1B	39.72	6.91	0.87	43.61	36.87	9.34	0.89	60.18
	Llama-3.2-3B	40.57	7.58	0.88	43.61	37.58	6.07	0.90	63.15
	Llama-3.1-8B	42.06	8.56	0.89	58.94	37.23	7.22	0.89	65.41
	Llama-2-13B	42.68	8.75	0.89	60.44	38.71	9.98	0.89	62.01
CLEF		Fine-tuning				Prompting			
		SARI↑	FKGL↓	BERT↑	LENS↑	SARI↑	FKGL↓	BERT↑	LENS↑
Baselines	Gold references	100.00	12.38	1.00	65.96	100.00	12.38	1.00	65.96
	MUSS_mined	33.55	10.96	0.91	48.38	33.55	10.96	0.91	48.38
	MUSS_wiki_mined	33.54	10.28	0.92	52.98	33.54	10.28	0.92	52.98
LLMs	Qwen2.5-1.5B	32.13	9.88	0.87	59.72	39.25	11.90	0.91	59.13
	Qwen2.5-7B	31.72	9.11	0.86	62.37	38.94	9.53	0.91	65.13
	Qwen2.5-14B	38.37	8.98	0.89	50.27	40.61	12.08	0.91	63.94
	Qwen2.5-32B	38.91	9.48	0.89	46.57	39.69	11.25	0.91	63.29
	DS-R1-1.5B	39.13	8.52	0.92	49.42	36.15	10.72	0.90	46.90
	DS-R1-7B	38.01	8.55	0.92	51.55	37.42	10.81	0.91	57.59
	DS-R1-14B	40.71	8.44	0.92	52.26	36.70	10.48	0.91	59.67
	DS-R1-32B	36.78	10.01	0.93	50.37	37.38	9.91	0.90	60.24
	Llama-3.2-1B	38.03	7.69	0.91	50.87	36.71	10.93	0.89	62.18
	Llama-3.2-3B	38.42	8.11	0.91	52.43	38.75	10.40	0.90	64.65
	Llama-3.1-8B	40.50	7.80	0.92	50.95	39.15	9.28	0.91	67.12
	Llama-2-13B	41.29	9.93	0.93	50.02	39.65	12.06	0.91	61.38
Newsela		Fine-tuning				Prompting			
		SARI↑	FKGL↓	BERT↑	LENS↑	SARI↑	FKGL↓	BERT↑	LENS↑
Baselines	Gold references	100.00	6.34	1.00	65.89	100.00	6.34	1.00	65.89
	MUSS_mined	35.30	9.50	0.91	61.20	35.30	9.50	0.91	61.20
	MUSS_wiki_mined	35.47	7.55	0.92	65.53	35.47	7.55	0.92	65.53
LLMs	Qwen2.5-1.5B	37.23	4.64	0.89	52.60	37.94	8.17	0.90	66.95
	Qwen2.5-7B	38.12	4.85	0.89	53.01	38.84	7.97	0.91	68.76
	Qwen2.5-14B	37.74	4.74	0.90	53.02	39.26	8.92	0.91	67.59
	Qwen2.5-32B	39.34	5.20	0.90	53.56	38.69	8.51	0.91	67.82
	DS-R1-1.5B	39.87	5.41	0.90	50.32	34.22	8.99	0.90	48.72
	DS-R1-7B	40.05	5.83	0.90	53.06	36.62	8.85	0.90	61.06
	DS-R1-14B	41.12	5.85	0.90	53.90	37.77	8.84	0.90	64.91
	DS-R1-32B	37.13	8.32	0.91	51.67	37.89	8.34	0.90	63.54
	Llama-3.2-1B	38.75	5.55	0.90	50.83	37.61	7.20	0.90	58.42
	Llama-3.2-3B	39.13	5.27	0.90	52.30	38.45	7.25	0.90	62.12
	Llama-3.1-8B	40.44	5.14	0.90	47.70	38.18	6.37	0.90	65.05
	Llama-2-13B	41.78	6.77	0.91	53.78	35.53	8.44	0.91	59.95

Table 6: Simplification Results on Med-Easi, CLEF, and Newsela.

Level 1		Fine-tuning				Prompting			
		SARI↑	FKGL↓	BERT↑	LENS↑	SARI↑	FKGL↓	BERT↑	LENS↑
Baselines	Gold references	100.00	8.77	1.00	62.48	100.00	8.77	1.00	62.48
	MUSS_mined	37.04	9.76	0.92	62.82	37.04	9.76	0.92	62.82
	MUSS_wiki_mined	37.93	7.75	0.93	66.72	37.93	7.75	0.93	66.72
LLMs	Qwen2.5-1.5B	32.88	6.28	0.89	50.94	36.92	10.01	0.93	63.71
	Qwen2.5-7B	31.49	5.88	0.88	52.46	37.75	9.29	0.92	66.83
	Qwen2.5-14B	32.66	6.02	0.89	51.66	38.91	9.90	0.92	64.51
	Qwen2.5-32B	38.80	7.83	0.91	52.52	37.13	9.32	0.92	67.60
	DS-R1-1.5B	39.04	7.57	0.92	48.67	35.52	9.25	0.91	50.52
	DS-R1-7B	38.78	8.04	0.92	52.13	36.06	9.04	0.91	62.43
	DS-R1-14B	41.41	8.59	0.92	52.82	36.74	9.67	0.92	64.66
	DS-R1-32B	35.88	9.21	0.93	50.67	37.13	9.19	0.92	60.81
	Llama-3.2-1B	37.20	8.64	0.91	45.74	36.57	9.02	0.92	53.49
	Llama-3.2-3B	35.98	7.23	0.89	50.45	36.33	9.11	0.91	58.63
	Llama-3.1-8B	38.56	6.44	0.89	31.68	35.88	6.36	0.91	58.46
	Llama-2-13B	40.68	9.39	0.93	49.82	35.76	8.68	0.91	64.27
Level 2		Fine-tuning				Prompting			
		SARI↑	FKGL↓	BERT↑	LENS↑	SARI↑	FKGL↓	BERT↑	LENS↑
Baselines	Gold references	100.00	7.16	1.00	65.29	100.00	7.16	1.00	65.29
	MUSS_mined	36.61	9.80	0.91	61.43	36.61	9.80	0.91	61.43
	MUSS_wiki_mined	36.90	7.82	0.92	65.73	36.90	7.82	0.92	65.73
LLMs	Qwen2.5-1.5B	37.96	5.32	0.90	53.76	37.83	8.29	0.90	67.70
	Qwen2.5-7B	39.66	5.95	0.90	53.73	38.89	8.38	0.91	67.64
	Qwen2.5-14B	38.35	5.82	0.90	54.85	38.34	9.64	0.91	66.54
	Qwen2.5-32B	39.90	5.76	0.90	54.97	38.56	8.70	0.91	68.07
	DS-R1-1.5B	39.33	6.65	0.90	51.09	35.23	9.27	0.90	48.96
	DS-R1-7B	39.26	6.02	0.90	52.41	37.38	9.05	0.90	61.18
	DS-R1-14B	40.72	6.56	0.90	53.73	37.58	8.88	0.90	65.00
	DS-R1-32B	38.71	8.78	0.90	53.04	37.86	8.35	0.90	65.20
	Llama-3.2-1B	38.63	6.03	0.90	52.49	36.38	7.18	0.90	57.18
	Llama-3.2-3B	39.22	6.32	0.90	53.65	38.80	6.61	0.91	56.30
	Llama-3.1-8B	40.56	6.34	0.90	53.14	37.59	6.64	0.90	67.84
	Llama-2-13B	41.60	7.91	0.91	53.45	36.19	8.97	0.92	54.58
Level 3		Fine-tuning				Prompting			
		SARI↑	FKGL↓	BERT↑	LENS↑	SARI↑	FKGL↓	BERT↑	LENS↑
Baselines	Gold references	100.00	5.41	1.00	67.82	100.00	5.41	1.00	67.82
	MUSS_mined	34.68	9.33	0.90	60.78	34.68	9.33	0.90	60.78
	MUSS_wiki_mined	34.97	7.43	0.91	65.17	34.97	7.43	0.91	65.17
LLMs	Qwen2.5-1.5B	38.89	4.13	0.89	54.83	38.10	7.75	0.89	66.60
	Qwen2.5-7B	40.19	4.00	0.89	52.84	39.55	7.10	0.90	71.60
	Qwen2.5-14B	39.95	4.19	0.89	54.03	39.67	8.25	0.90	68.35
	Qwen2.5-32B	38.02	4.18	0.89	54.65	39.88	7.83	0.90	69.09
	DS-R1-1.5B	40.56	4.40	0.90	51.67	33.91	8.79	0.90	48.08
	DS-R1-7B	40.65	5.26	0.90	53.28	37.17	8.74	0.90	60.79
	DS-R1-14B	41.65	4.62	0.90	55.36	38.69	8.51	0.90	65.41
	DS-R1-32B	35.53	7.55	0.90	50.91	38.43	7.97	0.90	64.16
	Llama-3.2-1B	39.62	4.28	0.89	53.25	37.33	6.42	0.89	63.84
	Llama-3.2-3B	40.47	4.35	0.90	53.53	39.19	6.14	0.90	69.24
	Llama-3.1-8B	41.48	4.47	0.90	53.53	39.46	6.31	0.90	67.32
	Llama-2-13B	42.40	5.51	0.90	55.63	35.49	7.97	0.90	58.29

Table 7: Simplification Results on Newsela levels (L1–L4) for Fine-tuning and Prompting.

Level 4		Fine-tuning				Prompting			
		SARI↑	FKGL↓	BERT↑	LENS↑	SARI↑	FKGL↓	BERT↑	LENS↑
Baselines	Gold references	100.00	4.03	1.00	67.96	100.00	4.03	1.00	67.96
	MUSS_mined	32.87	9.10	0.89	59.78	32.87	9.10	0.89	59.78
	MUSS_wiki_mined	32.06	7.19	0.90	64.51	32.06	7.19	0.90	64.51
LLMs	Qwen2.5-1.5B	39.17	2.81	0.89	50.87	38.92	6.64	0.89	69.80
	Qwen2.5-7B	41.14	3.57	0.89	53.02	39.16	7.09	0.89	68.96
	Qwen2.5-14B	40.00	2.91	0.90	51.53	40.11	7.88	0.89	70.94
	Qwen2.5-32B	40.62	3.01	0.89	52.10	39.18	8.18	0.89	66.52
	DS-R1-1.5B	40.54	3.00	0.89	49.85	32.21	8.65	0.89	47.30
	DS-R1-7B	41.49	4.00	0.89	54.41	35.85	8.58	0.89	59.85
	DS-R1-14B	40.71	3.64	0.89	53.70	38.06	8.31	0.89	64.57
	DS-R1-32B	38.40	7.74	0.89	52.04	38.14	7.85	0.89	64.00
	Llama-3.2-1B	39.54	3.24	0.89	51.83	39.02	6.16	0.89	59.16
	Llama-3.2-3B	40.84	3.17	0.89	51.58	39.46	7.13	0.89	64.30
	Llama-3.1-8B	41.17	3.30	0.89	52.43	39.78	6.18	0.89	66.59
	Llama-2-13B	42.44	4.25	0.90	56.22	34.68	8.14	0.90	62.65

Table 8: Simplification Results on Newsela Level 4 for Fine-tuning and Prompting (continued from Table 7).

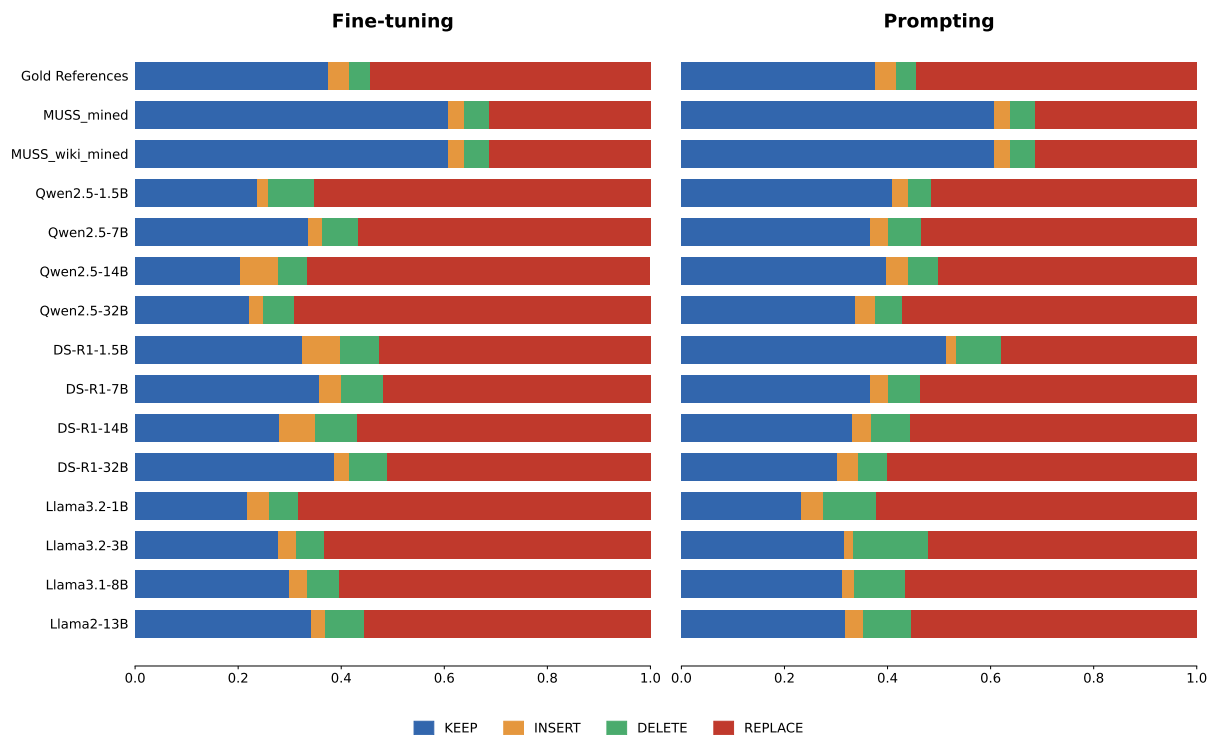


Figure 7: Distributions of Edit Operations on Med-Easi.

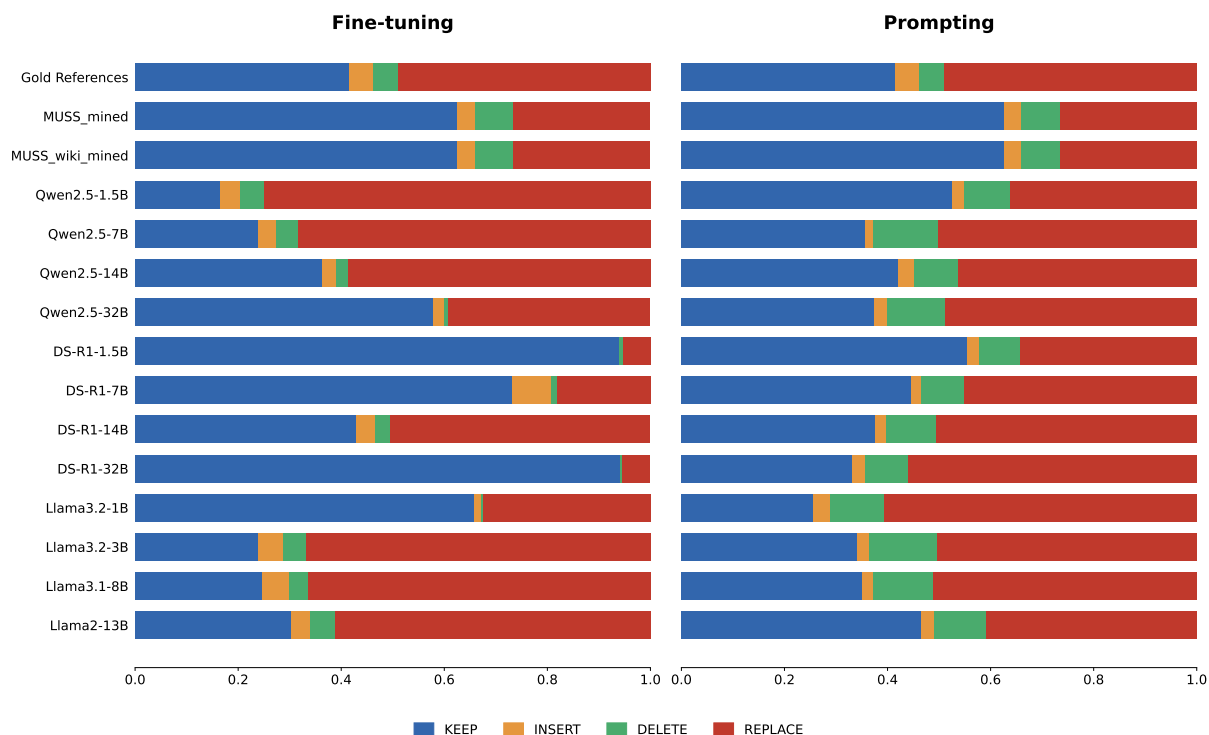


Figure 8: Distributions of Edit Operations on CLEF.

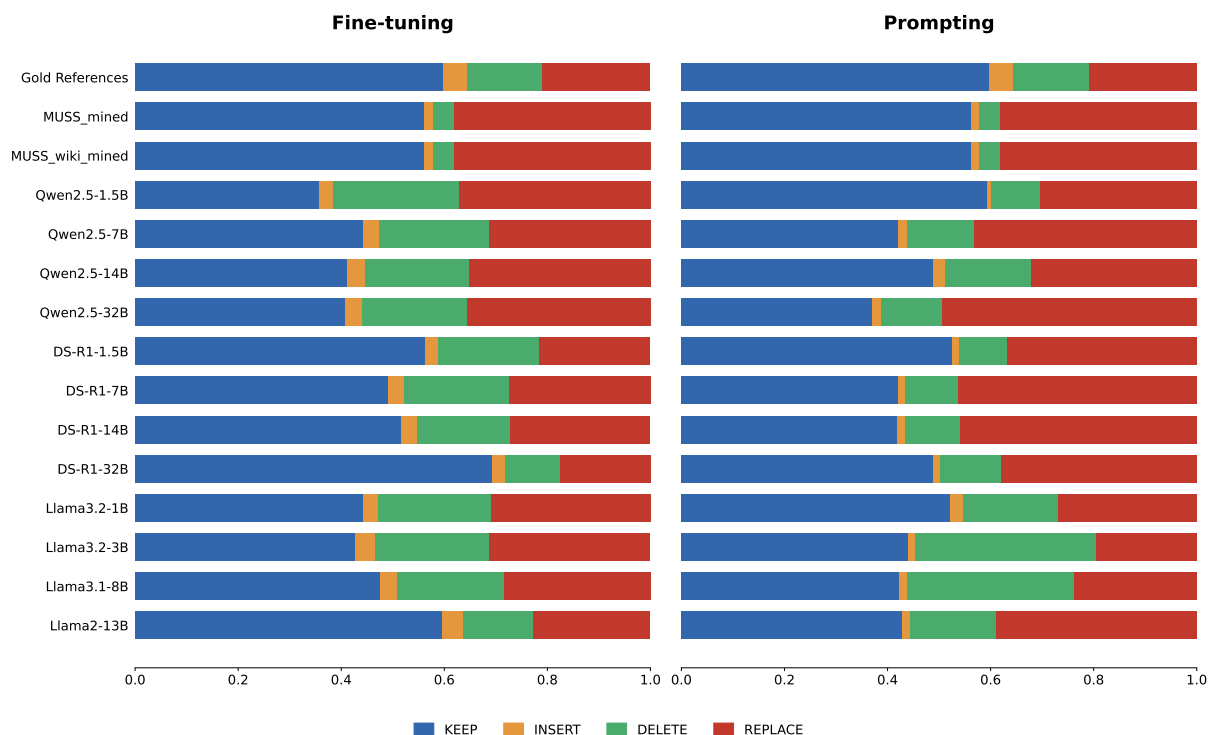


Figure 9: Distributions of Edit Operations on Newsela(Level 1).

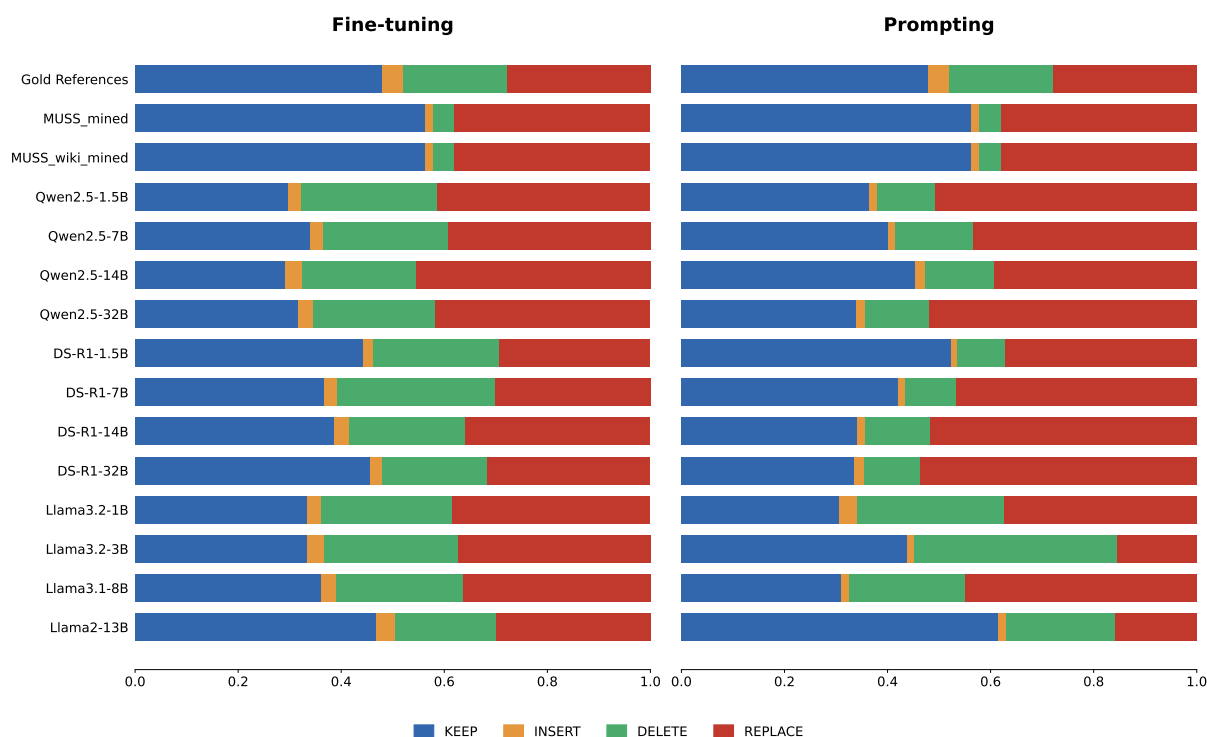


Figure 10: Distributions of Edit Operations on Newsela(Level 2).

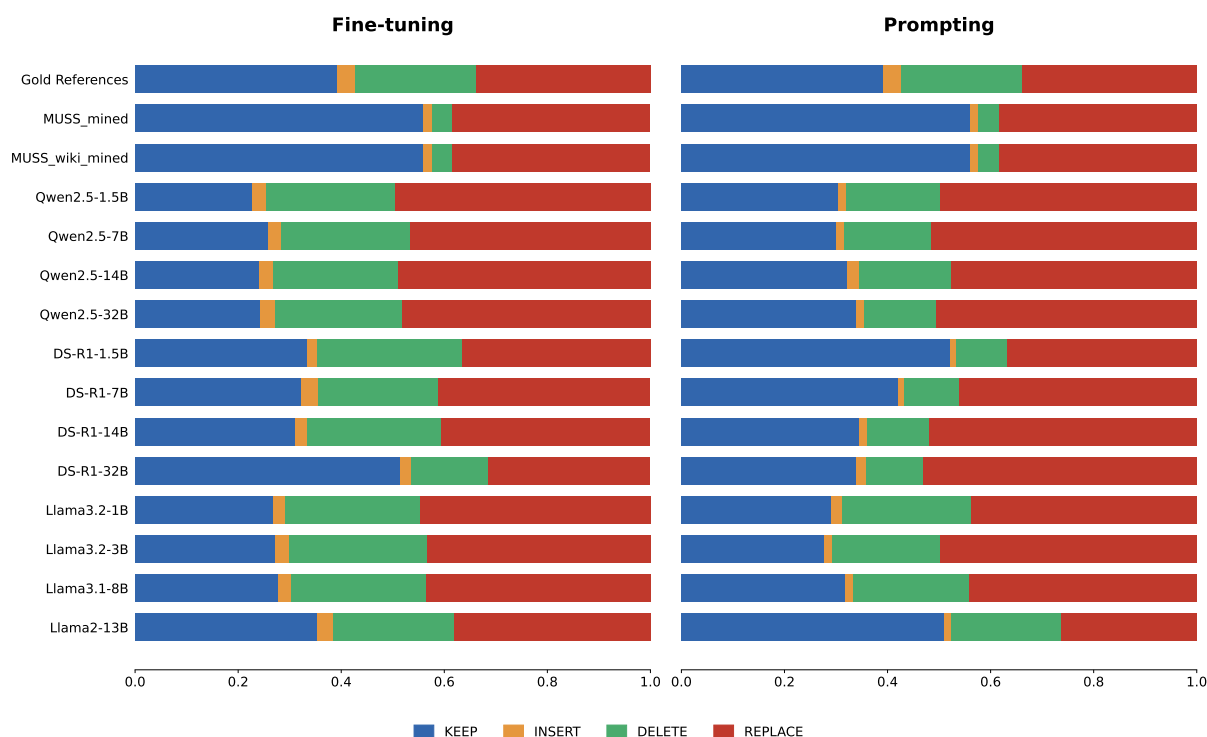


Figure 11: Distributions of Edit Operations on Newsela(Level 3).

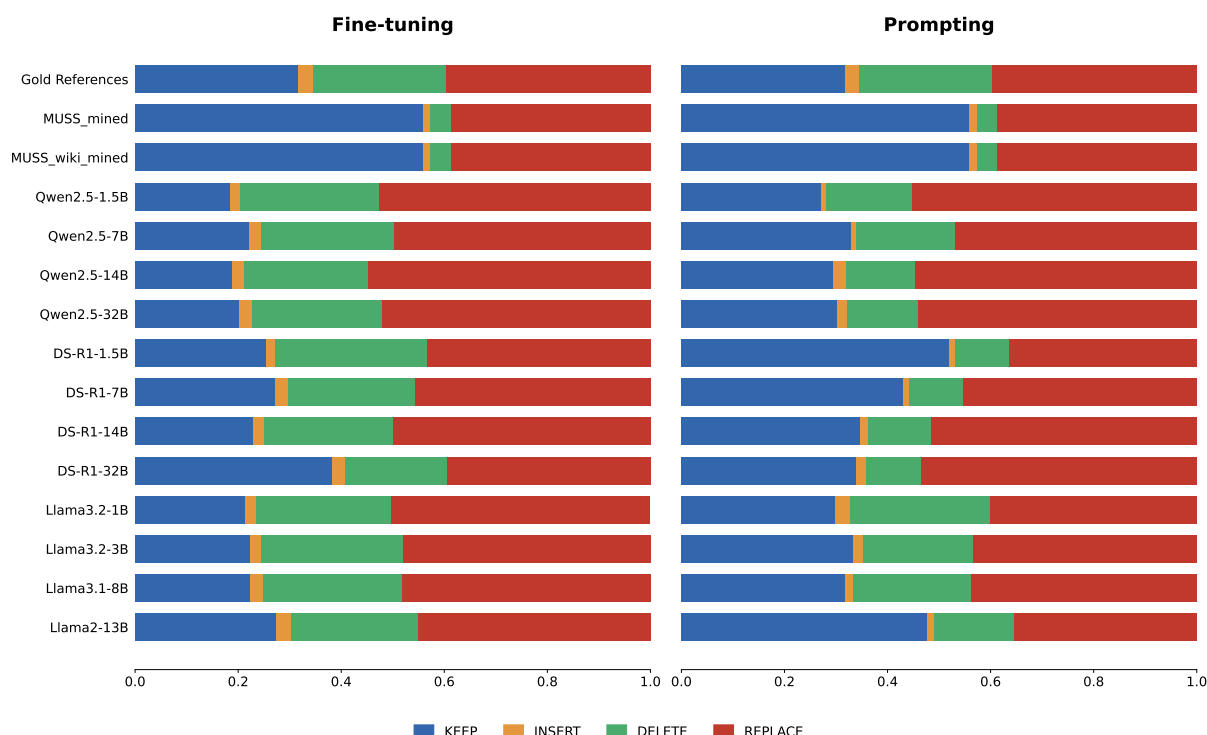


Figure 12: Distributions of Edit Operations on Newsela(Level 4).

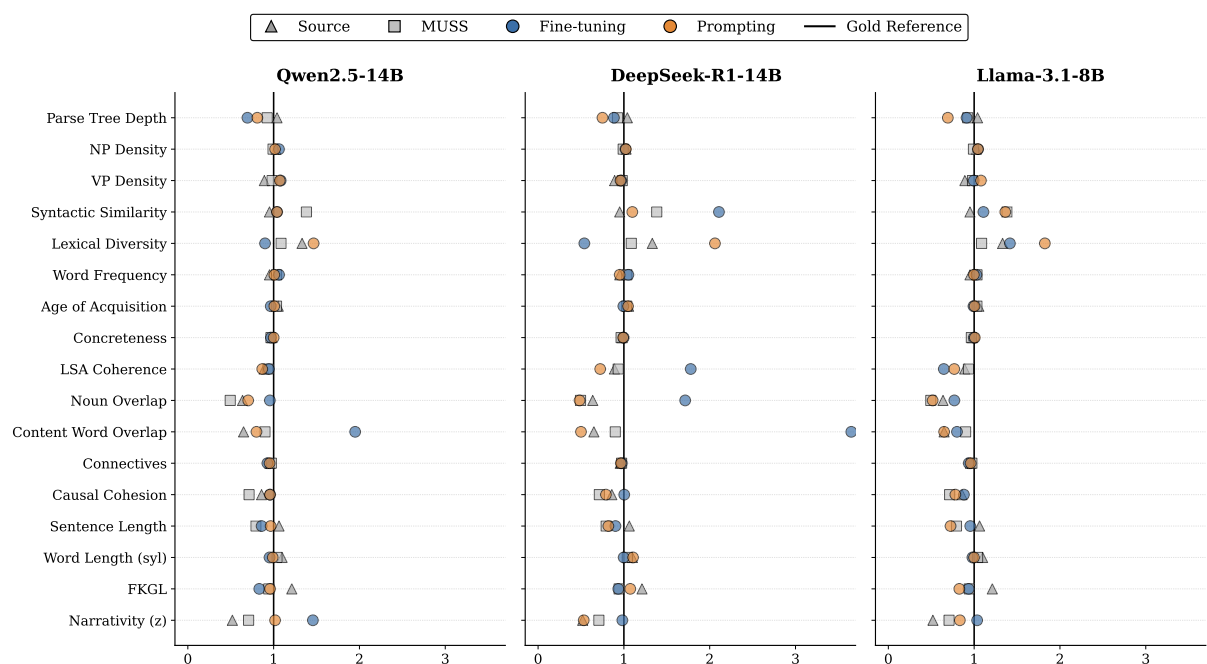


Figure 13: Linguistic feature profiles for the best-performing model per family (by LENS) on Med-Easi, normalized against the human reference (Rate = 1). Overall feature distributions are more concentrated than on ASSET (Figure 5), with deviations in syntactic similarity and narrativity largely absent. Lexical diversity is consistently higher under prompt engineering than fine-tuning across all three models, with the largest gap observed for Llama-3.1-8B. Content word overlap is generally elevated under fine-tuning, with DeepSeek-R1-14B showing a notably high ratio. Causal cohesion falls below gold-reference values for both methods, reflecting a systematic deficit in specialized domains.

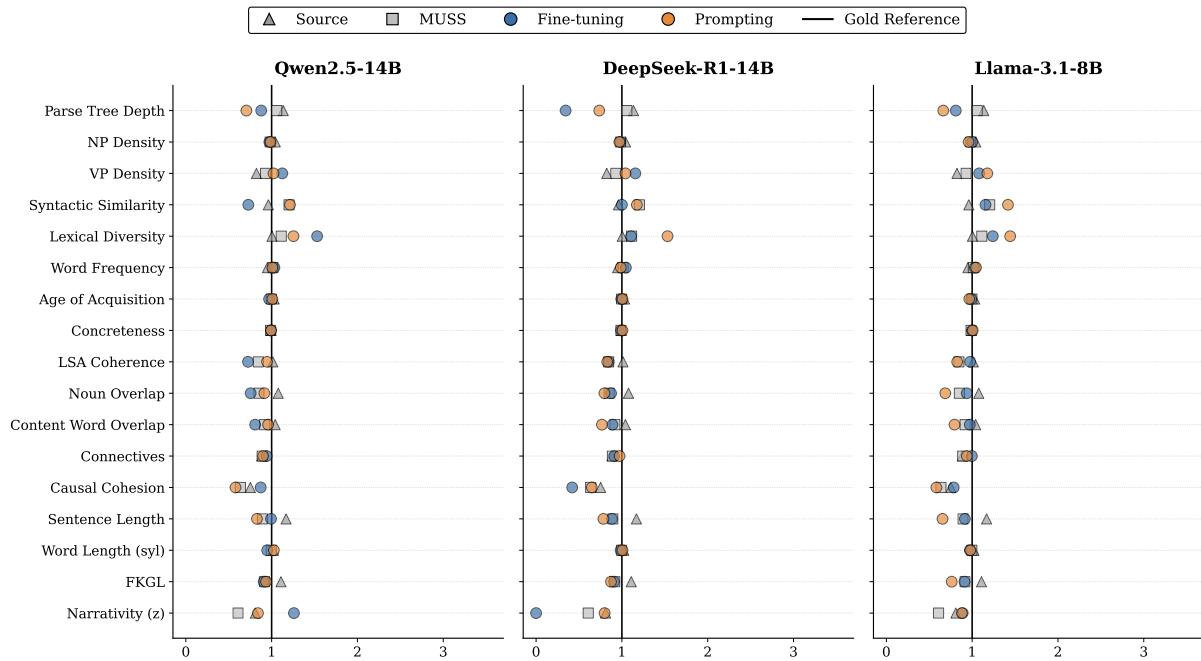


Figure 14: Linguistic feature profiles for the best-performing model per family (by LENS) on CLEF, normalized against the human reference (Rate = 1). Overall distributions are more concentrated than on ASSET, with deviations in syntactic similarity and narrativity largely absent. Lexical diversity is consistently higher under prompt engineering than fine-tuning, with the largest gap observed for Llama-3.1-8B. Content word overlap is generally elevated under fine-tuning, with DeepSeek-R1-14B showing a notably high ratio. Causal cohesion falls below gold-reference values for both methods, reflecting a systematic deficit in specialized domains.

Category	Linguistic Meaning	ID
Morpho-Syntactic	Parse Tree Depth	SYNLE
	NP Density	DRNP
	VP Density	DRVP
	Syntactic Similarity	SYNSTRUTa
Semantic-Lexical	Lexical Diversity	LDMTLD
	Word Frequency	WRDFRQc
	Age of Acquisition	WRDAOAc
	Concreteness	WRDCNCc
	LSA Coherence	LSASS1
Pragmatic-Discourse	Noun Overlap	CRFNOa
	Content Word Overlap	CRFCWOa
	Connectives	CNCAII
	Causal Cohesion	SMCAUSr
Style-Register	Sentence Length	DESSL
	Word Length	DESWLsy
	FKGL	RDFKGL
	Narrativity (z)	PCNARz

Table 9: Coh-Metrix linguistic feature indices and variable names.

Models	Sizes	Domain	Pre-training Data	Post-training Method
Qwen2.5	1.5B, 7B, 14B, 32B	General	Multilingual (18T)	SFT + DPO
DeepSeek-R1-Distill	1.5B, 7B, 14B, 32B	Reasoning	Based on Qwen2.5	SFT on reasoning traces
Llama-3	1B, 3B, 8B	General	Web, code, and multilingual (15T)	SFT + RS + DPO
Llama-2	13B	General	English-dominant (2T)	SFT + RS + RLHF (PPO)

Table 10: Summary of evaluated open-weight LLMs. All models are available in two variants: base models (used for fine-tuning) and instruction-tuned models (used for prompting). DeepSeek-R1-Distill models are initialized from Qwen2.5 base weights and fine-tuned on reasoning traces distilled from DeepSeek-R1.