Assessing Small Language Models as Text Simplification Evaluators

David Carranza Navarrete, Jan Bakker, Jaap Kamps
Institute for Logic, Language and Computation (ILLC)
University of Amsterdam
Amsterdam, The Netherlands
david.carranza.navarrete@student.uva.nl, j.bakker@uva.nl, kamps@uva.nl

Abstract

Text simplification requires reliable automatic evaluation, yet existing learnable metrics such as LENS and LENS-SALSA are specialized and costly to develop. Moreover, it remains unclear how these metrics compare to using large language models (LLMs) as evaluators. Exploring this question is important because LLM-based evaluation could make simplification research and deployment more flexible and easier to adapt than training new task-specific metrics for each setting. In this work, we empirically compare several small, open-weight instruction-tuned LLMs with LENS and LENS-SALSA in both reference-based and reference-free evaluation settings. We measure their alignment with human judgments across multiple datasets. Our results provide insight into when small LLMs can serve as effective evaluators and when specialized metrics remain preferable, informing the design of future evaluation pipelines for text simplification and related text generation tasks.

Lay Summary: *The evaluation of text simplification output remains a great challenge in terms of building reusable evaluation corpora and evaluation measures. Traditional reference-based evaluations are imprecise as references only cover one or a few possible simplifications. Learned measures still require extensive labeled data for training, and may not generalize to new domains. LLM-based evaluation presents a pragmatic alternative in case no extensive references are available. In this paper, we systematically compare these evaluation approaches against human ratings.*

Keywords: LENS, Automatic Evaluation, LLM-as-a-Judge

1. Introduction

Text simplification aims to make content more accessible while preserving its original meaning. Reliable evaluation of simplification quality is therefore essential for both system development and deployment. Traditionally, evaluation relies on human judgments, which are expensive and time-consuming. To address this, automatic metrics such as LENS (Maddela et al., 2023) and its reference-free variant LENS-SALSA (Heineman et al., 2023) have been proposed as learnable metrics for text simplification. These metrics attempt to approximate human judgments using supervised models trained on annotated data.

Recent advances in large language models (LLMs) have used generative models as evaluators (Gao et al., 2025). Instead of relying on task-specific learned metrics, LLMs are prompted to act as judges and directly assign quality scores. This raises an important question: how well do relatively small instruction-tuned LLMs perform as judges of simplification quality compared to specialized learnable metrics like LENS?

In this work, we investigate whether small, open-weight LLMs can serve as reliable simplification judges. We focus on three popular open-source models: Microsoft's `Phi-3-mini-4k-Instruct`, Alibaba Cloud's `Qwen2.5-7B-Instruct`, and Meta's `Llama-3.1-8B-Instruct`. Specifically, we investigate LLM's performance against LENS in both reference-based and reference-free settings. We evaluate: (1) LENS-SALSA (metric without references) against LLM judges without references and (2) LENS (metric with references) against LLM judges with access to references. Finally, we compute correlations between LLM-assigned scores and human judgments to assess alignment with human evaluation.

2. LLMs as Judges

We investigate whether small instruction-tuned large language models (LLMs) can serve as automatic judges of text simplification quality. Unlike supervised evaluation metrics that are explicitly trained for simplification assessment, these models perform evaluation through prompting at inference time. Specifically, the models are asked to assess key aspects of simplification quality, including fluency, meaning preservation, and simplicity.

We experiment with the following open-weight instruction-tuned LLMs:

- `Phi-3-mini-4k-Instruct`[1]

[1] https://huggingface.co/microsoft/Phi-3-mini-4k-instruct

Proceedings of the 2nd Workshop on Evaluating Text Difficulty in a Multilingual Context (DeTermIt! 2026), pages 83–86

- `Qwen2.5-7B-Instruct`[2]
- `Llama-3.1-8B-Instruct`[3]

These models range from approximately 3.8B to 8B parameters and were selected to represent compact open-weight models that remain practical for research and deployment settings.

3. Methodology

Non-learnable evaluation metrics are commonly used to assess text without additional training. For example, BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002) measures n-gram overlap between candidate and reference texts, focusing on precision but lacking semantic sensitivity. SARI (System Output Against References and against the Input sentence) (Xu et al., 2016) evaluates the quality of added, deleted, and retained words, better capturing meaning preservation and simplicity. BERTScore (Zhang et al., 2019) uses contextual embeddings to measure semantic similarity, making it more robust to paraphrasing. Readability metrics such as FKGL (Flesch–Kincaid readability tests) (Kincaid et al., 1975) estimate the education level required to understand the text.

In this work, we use two evaluation settings to compare LLM judges with existing learnable and non-learnable automatic metrics. We use the LENS[4] and LENS-SALSA[5] checkpoints released on HuggingFace to evaluate simplification performance using the top-k outputs, reporting results for $k = 3$. We use the corresponding code[6] for our experiments and present our results in the same manner as Maddela et al. (2023), additionally including results from our LLM-based models.

Reference-free. In the reference-free setting, the model receives the original complex sentence and the simplified candidate. The model is prompted to assess the quality of the simplification by assigning scores based on three criteria: meaning preservation, fluency, and simplicity. This setup mirrors the reference-free evaluation paradigm used by metrics such as LENS-SALSA.

Reference-based. In the reference-based setting, the model receives the original complex sentence, the simplified candidate, and one or more human-written reference simplifications. The model is asked to evaluate the quality of the candidate simplification while considering the reference simplifications as guidance for expected outputs. This setup corresponds to traditional reference-based evaluation used by metrics such as LENS.

3.1. Scoring and Correlation

Each setting produces scalar quality scores. We compute Pearson correlation for datasets with continuous human ratings and Kendall τ-like correlation for ranking-based datasets. Alignment with human judgments is used as the primary measure of evaluation quality.

4. Experiments

4.1. Datasets

We evaluate across three benchmark datasets:

- **SimpEval**$_{2022}$: Human rankings of simplifications evaluated using a Kendall τ-like correlation. (Maddela et al., 2023)
- **Wiki-DA**: Direct assessment scores for fluency, meaning, and simplicity. (Alva-Manchego et al., 2021)
- **Newsela-LIKERT**: Human Likert-scale ratings across grammaticality, meaning preservation, and simplicity. (Maddela et al., 2021)

These datasets cover both ranking-based and direct scoring evaluation paradigms.

4.2. Evaluation Settings

We compare:

1. Reference-free LLM judges vs LENS-SALSA.
2. Reference-based LLM judges vs LENS.

All correlations are computed against human ratings provided in the respective datasets.

5. Results

Table 1 presents the correlation between automatic metrics and human judgments in the reference-based setting. LENS achieves the highest correlations overall, particularly on Wiki-DA and Newsela-LIKERT. However, several LLM judges obtain competitive results on SimpEval and Wiki-DA, suggesting that instruction-tuned models can approximate dedicated evaluation metrics when provided with references.

Table 2 reports the reference-free evaluation results. LENS-SALSA shows strong performance

[2] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

[3] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[4] https://huggingface.co/davidheineman/lens

[5] https://huggingface.co/davidheineman/lens-salsa

[6] https://github.com/Yao-Dou/LENS

	SimpEval$_{2022}$			Wiki-DA			Newsela-Likert		
Metric / Model	Para	Split	All	Fluency	Meaning	Simplicity	Fluency	Meaning	Simplicity
FKGL	-0.397	-0.318	-0.331	0.084	0.185	0.037	0.169	0.293	-0.053
BLEU	0.048	-0.054	-0.033	0.460	0.622	0.438	0.333	0.261	0.121
SARI	0.206	0.140	0.149	0.335	0.534	0.366	0.234	0.122	0.101
BERTScore	0.238	0.085	0.106	0.642	0.699	0.622	0.389	0.295	0.206
LENS (*k*=3)	**0.429**	0.333	0.331	**0.807**	0.660	**0.750**	**0.621**	**0.431**	**0.362**
LLM Judges (with references)									
Qwen2.5-7B-Instruct	**0.818**	**0.358**	**0.457**	0.630	0.716	0.680	0.458	0.393	0.265
Llama-3.1-8B-Instruct	0.333	0.320	0.358	0.618	**0.761**	0.639	0.379	0.313	0.208
Phi-3-mini-4k-Instruct	0.130	0.347	0.325	0.605	0.689	0.639	0.506	0.371	0.259

Table 1: Comparison between traditional automatic metrics from the LENS paper and LLM judges (with references). Pearson correlations with human ratings are reported. Higher values indicate better alignment with human evaluation. Best results are in bold and second best are underlined.

	SimpEval$_{2022}$			Wiki-DA			Newsela-Likert		
Metric / LLM	Para	Split	All	Fluency	Meaning	Simplicity	Fluency	Meaning	Simplicity
LENS-SALSA	0.263	0.212	0.229	0.701	0.676	0.640	0.497	0.356	0.284
Qwen2.5-7B-Instruct	-0.353	0.117	-0.028	**0.648**	**0.802**	**0.682**	**0.510**	**0.572**	**0.272**
Llama-3.1-8B-Instruct	**0.333**	**0.500**	**0.083**	0.586	0.740	0.620	0.428	0.520	0.233
Phi-3-mini-4k-instruct (**)	0.000	0.000	0.000	0.021	0.026	0.051	0.000	0.000	0.000

Table 2: Correlation between automatic evaluators and human judgments across simplification datasets. LENS-SALSA is a dedicated reference-free metric. LLM judges operate without references. Best values are in bold; second best are underlined.

across datasets, while LLM judges demonstrate varying levels of correlation with human judgments. These findings suggest that LLM-based evaluation may provide a flexible alternative to specialized metrics, although performance depends on both the dataset and evaluation dimension.

6. Conclusion

In this work, we investigated whether small instruction-tuned large language models can serve as automatic evaluators for text simplification. We compared three open-weight LLM judges with dedicated simplification evaluation metrics, including LENS and its reference-free variant LENS-SALSA, across three benchmark datasets.

The experiments show that LLM judges can achieve moderate correlations with human judgments and, in some cases, approach the performance of traditional metrics, particularly in reference-based evaluation settings. However, specialized metrics such as LENS and LENS-SALSA remain more consistent and generally achieve stronger correlations with human evaluation.

These findings suggest that while small LLMs can provide flexible and lightweight evaluation signals, dedicated metrics still offer advantages in reliability and stability. Future work could explore improved prompting strategies, calibration methods, or hybrid approaches that combine LLM-based evaluation with learned metrics to further improve automatic evaluation for text simplification.

7. Limitations and future work

This study focuses on sentence-level simplification in English but does not address document-level evaluation or other languages. Additionally, we restrict our analysis to relatively small open-weight LLMs and do not compare against larger proprietary models. Future work could explore strategies for LLM judges, pairwise preference prompting, or training smaller evaluation models using LLM-generated supervision.

Acknowledgments

Jan Bakker and Jaap Kamps are supported by the Netherlands Organization for Scientific Research (NWO NWA # 1518.22.105). Jaap Kamps is also supported by the University of Amsterdam (AI4FinTech program) and ICAI (AI for Open Government Lab). Views expressed in this paper are

not necessarily shared or endorsed by those funding the research.

8. Bibliographical References

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.

Mingqi Gao, Xinyu Hu, Xunjian Yin, Jie Ruan, Xiao Pu, and Xiaojun Wan. 2025. LLM-based NLG evaluation: Current status and challenges. *Computational Linguistics*, 51:661–687.

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, Singapore. 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. Research Branch Report 8-75.

Mounica Maddela, Fernando Alva-Manchego, and Wei Xu. 2021. Controllable text simplification with explicit paraphrasing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3536–3553, Online. Association for Computational Linguistics.

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

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 (ACL)*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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

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

A. Appendices

A.1. LLM Evaluation Prompts

The following prompts were used to instruct the language models to evaluate the quality of simplified sentences. The Reference-Free prompt is based on the instructions given to human annotators to rate simplifications by (Maddela et al., 2023). The Reference-Based makes use of the simplification, the original text, and the reference to grade.

LLM Evaluation Prompt (Reference-Free)

```
You are an evaluator specialized in text simplification that
scores how well a simplified sentence improves on an original
sentence.
Assign an integer score from 0–100:
100 – Fully simplified, fluent, preserves core meaning
75 – Somewhat simpler, mostly fluent, meaning close
50 – Simpler, somewhat fluent, meaning similar
25 – Equally simple, some fluency, meaning lost
0 – Completely unreadable
Higher scores indicate better meaning preservation, fluency,
and simpler wording.
Return:
1) "score: X" (0–100)
2) Short explanation (1–2 sentences)
Original: """<original>"""
Simplified: """<simplified>"""
```

LLM Evaluation Prompt (Reference-Based)

```
You are an evaluator for text simplification. Rate how well
the simplification improves on the original.
When a REFERENCE is provided, treat it as a gold-standard
simplification and compare the candidate to both the original
and the reference.
Scoring rules:
100 – Fully simplified, fluent, preserves core meaning and
aligns with reference
75 – Somewhat simpler, mostly fluent, meaning close
50 – Simpler, somewhat fluent, meaning similar
25 – Equally simple, some fluency, meaning lost
0 – Completely unreadable
Return:
1) "score: X" (0–100)
2) Short explanation
Original: """<original>"""
Reference: """<reference>"""
Simplified: """<simplified>"""
```