Evaluation Under Imperfect Benchmarks and Ratings: A Case Study in Text Simplification

Anonymous authors
Paper under double-blind review

Abstract

Despite the successes of language models, their evaluation remains a daunting challenge for new and existing tasks. We consider the task of text simplification, commonly used to improve information accessibility, where evaluation faces two major challenges. First, the data in existing benchmarks might not reflect the capabilities of current language models on the task, often containing disfluent, incoherent, or simplistic examples. Second, existing human ratings associated with the benchmarks often contain a high degree of disagreement, resulting in inconsistent ratings; nevertheless, existing metrics still have to show higher correlations with these imperfect ratings. As a result, evaluation for the task is not reliable and does not reflect expected trends (e.g., more powerful models being assigned higher scores). We address these challenges for the task of text simplification through three contributions. First, we introduce SynthSimpliEval, a synthetic benchmark for text simplification featuring simplified sentences generated by models of varying sizes. Through a pilot study, we show that human ratings on our benchmark exhibit high inter-annotator agreement and reflect the expected trend: larger models produce higher-quality simplifications. Second, we show that auto-evaluation with a panel of LLM judges (LLMs-as-a-Jury) often suffices to obtain consistent ratings for the evaluation of text simplification. Third, we demonstrate that existing learnable metrics for text simplification benefit from training on our LLMs-as-a-Jury-rated synthetic data, closing the gap with pure LLMs-as-a-Jury for evaluation. Overall, through our case study on text simplification, we show that a reliable evaluation requires higher quality test data, which could be obtained through synthetic data and LLMs-as-a-Jury ratings.

1 Introduction

Despite advances in LLMs, evaluating the quality of their generations remains a challenge Pillutla et al. (2021); Chang et al. (2023). One such task is text simplification: crucial to improving the clarity and accessibility of information, making content easier to understand for a wider audience (Al-Thanyyan & Azmi, 2021). The gold standard for text simplification evaluation is human judgment, which provides the most direct and comprehensive assessment of simplification quality (Devaraj et al., 2022; Maddela et al., 2023). However, existing human evaluation of text simplification can be unreliable due to low agreement among annotators on simplification ratings (Wu & Arase, 2024; Popović et al., 2022). This makes it difficult to establish a universally reliable evaluation standard.

To address this challenge, we propose a novel text simplification benchmark named SynthSimpliEval featuring a dataset of complex sentences and their corresponding simplifications generated by LLMs of varying sizes. Through a pilot study with new human annotators, we observe high inter-annotator agreement on our benchmark. Moreover, the human study shows a strong correlation between the score assigned to each simplification and the size of the model that generated it. This finding is consistent with prior research demonstrating that, within the same model family, larger models consistently outperform smaller ones across a range of NLP tasks (Hestness et al., 2017; Kaplan et al., 2020; Hoffmann et al., 2022; Liang et al., 2023; McKenzie et al., 2023).

To alleviate the cost of human labeling, we subsequently adopt an LLMs-as-a-Jury approach (Verga et al., 2024; Chan et al., 2023; Wang et al., 2024) to evaluate simplification quality, aggregating the scores across all models in the jury to obtain the final score. We conduct a systematic ablations study on how various design choices, such as prompting strategy, rationale inclusion, and aggregation method, affect the assigned simplification score. We find that few-shot prompting with rationale generation, combined with score averaging, yields simplification scores that best correlate with model sizes. Moreover, we compare inter-LLM agreement for unified scoring against the multi-dimensional approach used in previous benchmarks (Wubben et al., 2012; Maddela et al., 2023; Alva-Manchego et al., 2020), and find that unified scoring results in higher agreement. This suggests that unified scoring is both simpler to apply and more reliable as an evaluation method.

Motivated by the observed trend that larger models produce higher-quality simplifications, we evaluate the performance of existing text simplification metrics (Flesch, 1948; Maddela et al., 2023; Zhang et al., 2019; Cripwell et al., 2023) and LLMs-as-a-Jury on SynthSimpliEval by measuring the correlation between assigned scores and model sizes. Existing metrics struggle to consistently reflect the trend, whereas LLMs-as-a-Jury reliably assigns higher scores to outputs from larger models, aligning with expectations on simplification quality.

In order to improve existing learnable metrics using our synthetic data, we use the same approach to generate synthetic complex sentences and create a set of LLM-rated complex-simple sentence pairs. We train a small neural network on sentence embeddings of these scores, following previous work (Maddela et al., 2023; Huang & Kochmar, 2024). We find that with this new set of training data, the network recovers some of the capability of the LLMs, achieving a higher correlation with model size than all previous methods.

Overall, through our case study on text simplification, we show that a reliable evaluation requires high-quality data. By leveraging synthetic benchmarks in place of traditional human annotations, we are able to reliably evaluate text simplification metrics. This approach offers a practical recipe for evaluation in other tasks where high-quality annotated data may be limited or unreliable. We will publicly release our code and data.

2 Existing Text Simplification Benchmarks and Ratings

When evaluating text simplification metrics, existing work often relies on datasets that contain both complex-simple sentence pairs and their corresponding human ratings (Xu et al., 2015; 2016; Maddela et al., 2023). Formally, given a source sentence c, a target simplification t, and optionally, a set of reference (human) ratings $r_1(c,t), ..., r_n(c,t)$, the task of evaluating text simplification is to compute a real-valued score $q(c,t)$. The strength of the evaluation is usually measured by the correlation of q with the (aggregated) reference ratings. Evaluation methods are considered reliable if they have high correlation with human raters. Hence, it is very important to consider the quality of source complex sentences c and ratings r in a benchmark for reliable evaluation.

We investigate three such benchmarks commonly used for text simplification: Simplicity-DA (Alva-Manchego et al., 2021), Newsela-Likert (Maddela et al., 2021), and SimpEval2022 (Maddela et al., 2023). The Simplicity-DA dataset, also referred to as WikiDA, consists of 600 simplifications of 100 complex sentences collected from Wikipedia. The simplifications are sourced from the TurkCorpus test set (Xu et al., 2016) and are produced by six older systems: PBMT-R, Hybrid, SBMT-SARI, Dress-Ls, DMASS-DCSS, and ACCESS. These are rated by humans on three separate criteria (fluency, meaning, and simplicity) on a scale of 0-100. The Newsela-Likert dataset contains 500 simplifications from 100 complex sentences sourced from the Newsela dataset (Xu et al., 2015), with one human simplification and four system simplifications. The systems are BERT-Initialized Transformer (Jiang et al., 2020), EditNTS (Dong et al., 2019), LSTM (Zhang & Lapata, 2017), and Hybrid-NG (Narayan & Gardent, 2014). Newsela-Likert also contains human ratings on fluency, meaning, and simplicity, but on a five point scale. Lastly, SimpEval2022 contains 360 simplifications from 60 complex sentences, collected from Wikipedia and simplified by humans and previously SOTA models including GPT-3.5 (Brown, 2020) and T5 (Raffel et al., 2020). SimpEval2022 is also scored on a 0-100 scale.

Source	Complex Sentence	Simplified Sentence
Simplicity-DA	For example, the stylebook of the Associated Press is updated annually.	, the stylebook is updated.
Newsela-Likert	companies want to drill for oil in the park.	companies want to drill for oil.
SimpEval2022	On the fifth day of flight, November 20, 2022, at 19:09 UTC, the Orion spacecraft entered the Lunar sphere of influence, thus the Moon's gravitational force became stronger than Earth's relative to the spacecraft.	On the fifth day of flight, November 20, 2022, at 19:09 UTC, the Orion spacecraft entered the Lunar sphere of influence, so the Moon's gravitational force became stronger than Earth's relative to the spacecraft.
SynthSimpliEval	The study's findings suggest a significant correlation between long-term cannabis use and alterations in neurocognitive function, particularly in attention and memory processes.	Using cannabis for a long time might change how your brain works, especially when it comes to paying attention and remembering things.

Table 1: Example sentences from various datasets. The first complex sentence has a simplification that few modern systems would output as it is grammatically incorrect. Newsela-Likert and SimpEval2022 themselves are not challenging enough for modern LLMs to produce a useful simplification. Additional samples can be found in Appendix E.

In our analysis, we find two major problems across existing datasets. First, many of the sentence pairs are overly simplistic or incoherent; examples are shown in Table 1. As such, they might not reflect modern systems or accurately assess the model's simplification capabilities. Second, we find lower agreement among human raters, indicating subjectivity of the annotation task, possible underspecification of the task instructions, and issues with annotation collection (Nowak & Rüger, 2010; Aroyo & Welty, 2014; Hsueh et al., 2009). Coupled with the relatively small sample size for each sentence pair, it is difficult to construe these ratings as a reliable gold standard.

2.1 Coherence and Difficulty Gaps in Existing Datasets

We first investigate the sentence pairs in existing datasets and find many that are not reflective of the output of modern simplification systems such as LLMs. This is particularly prevalent in older datasets, where simplifications were often generated by dated systems as opposed to written by hand. A few example simplifications are shown in Table 1. Many low-quality simplifications in existing datasets (Alva-Manchego et al., 2021) reflect failure modes that modern LLMs rarely exhibit, like grammatical errors and disfluencies (Reinhart et al., 2025). This results in a mismatch between the kinds of "bad" simplifications present in these datasets and the kinds of errors made by contemporary systems, highlighting the need for new, more representative examples of low-quality simplifications.

Secondly, the overall difficulty level of existing datasets is relatively low (Maddela et al., 2021; 2023) and does not reflect the complexity of potential real-world applications, such as those in medicine or law. Sentences that are considered difficult in existing datasets often require moderate in-domain knowledge to understand, but the vocabulary itself is not overtly technical nor the sentence structure grammatically complex. While this is not inherently problematic, as simplification can be evaluated even on simple inputs, the lack of linguistic and semantic complexity means these datasets fail to adequately challenge modern LLMs. As LLMs can now reliably produce fluent and coherent simplifications (§3.1), such data does little to test their capabilities in preserving nuance, accurately simplifying domain-specific content, or avoiding hallucinations (Devaraj et al., 2022), falling short of providing a robust evaluation for current systems.

Dataset	Human Fluency	Human Meaning	Human Simplicity	LLM Fluency	LLM Meaning	LLM Simplicity	Human Unified	LLM Unified
Newsela	0.153	0.213	0.054	0.672	0.650	0.640	—	0.657
SimpEval	—	—	—	—	—	—	0.228	0.320
SynthSimpliEval	—	—	—	—	—	—	0.671*	0.619

Table 2: Intraclass Correlation Coefficient, ICC(2, 1) for human and LLM ratings on three benchmarks. Human unified ratings(*) are on a subset of SynthSimpliEval (see §3.2).

2.2 Annotator Disagreement

Previous literature has shown that human annotators are not always in agreement across many other NLP tasks (Castilho, 2020). In our preliminary experiments, we find that this pattern holds in existing text simplification datasets. We compute inter-annotator agreement across the Newsela and SimpEval2022 datasets and report the $ICC(2,1)$ (Intraclass Correlation) scores in Table 2. ICC is often used to compute consistency across raters on non-categorical data (Hackl et al., 2023). We find that individual raters are on average unreliable, which brings the relevance of their averages into question. It may still be possible to find a more accurate rating through larger sample sizes, but this is often prohibitively expensive.

Moreover, text simplification a relatively subjective task and individual performance heavily depends on the quality of instructions and examples provided. What is "simple" can vary significantly depending on the educational background, the familiarity of the topic, and prior knowledge (Aroyo & Welty, 2015; Liu et al., 2018). Previous literature does not report the educational background of its annotators, nor does it attempt to control for variability across annotator profiles (Snow et al., 2008). A correct-sounding, but semantically incorrect, simplification of a technical statement may not be correctly identified by a rater unfamiliar with the subject. Secondly, it is important for the raters to have a strong set of guidelines in performing the ratings. Previous work often leaves this up to interpretation, allowing the rater to decide on their own definition of "simplicity" (Alva-Manchego et al., 2021). This lack of transparency and standardization adds another layer of uncertainty to the reliability of the gold standard ratings in the datasets.

3 SynthSimpliEval: A Synthetic Benchmark for Evaluating Text Simplification

To address the challenges in existing benchmarks and corresponding human ratings, we introduce a new synthetic benchmark, or SynthSimpliEval, using Qwen 2.5 72B Instruct (§3.1). We assess the quality of our benchmark on the basis of two criteria (§3.2). First, we measure the agreement among human annotators for simplifications in a pilot study. Second, we report the performance of models of different sizes, based on our human pilot.

3.1 Constructing SynthSimpliEval

We construct our dataset by combining human-written and synthetic complex sentences. We use all 60 news article sentences from SimpEval2022, which are more challenging than other datasets, with an average FKGL of 18.29 compared to 8.79 for Newsela-Likert and 10.61 for Simplicity-DA. Additionally, we generate 200 synthetic sentences using Qwen 2.5 72B Instruct (Qwen et al., 2025). This was designed with 20 knowledge domains and 750 concept nouns (full lists in Appendix D), resulting in a diverse and challenging (average FKGL of 19.51) test set. See example sentences in Table 3 and the prompt used in Appendix C.

We collect simplifications of 260 complex sentences from four instruction-tuned Llama 3 models (1B, 3B, 8B, and 70B Instruct) using the same 2-shot prompt (details in Appendix C). With a total of 1040 simplified sentences whose relative quality we compare in Table 4, our dataset is comparable or slightly larger than prior works.

Domain and Concept Noun	Generated Sentence
architecture grocery	The cantilevered roofline of the facility must accommodate 30-degree angular deviations in structural supports while maintaining a 3-inch minimum clearance from refrigerated storage units.
mathematics vacation	The optimization of vacation scheduling for a group can be modeled as a constraint satisfaction problem, where the objective function minimizes the total dissatisfaction across all participants, subject to constraints on available dates and group size.
anthropology alcohol	Ethnographic studies reveal that the ritualized consumption of alcohol in social gatherings serves as a significant mediator of interpersonal relationships and community cohesion in diverse cultural contexts.

Table 3: Synthetic complex sentence samples from SynthSimpliEval.

In our benchmark, we use model size as a direct proxy for simplification quality, motivated by prior findings that LLMs generally produce higher-quality outputs (§1). We operationalize this intuition by selecting four LLaMA 3 models of increasing size (1B, 3B, 8B, and 70B), assuming larger models generate better simplifications. Rather than assigning numerical scores, we evaluate metrics based on their ability to correctly rank simplifications by model size using Spearman correlation, a method validated by our subsequent human study.

3.2 Human Evaluation: Correlation with Model Size and Annotator Agreement

To validate our assumption that larger models produce better simplifications, we conduct a human evaluation on 20 randomly selected complex sentences from SynthSimpliEval (10 each from SimpEval2022 and our synthetic dataset), paired with simplifications from four models (totaling 80 pairs). Three NLP-expert annotators rate simplification quality on a 5-point Likert scale using consistent instructions (Appendix C.2). As shown in Table 2, Inter-annotator agreement measured in ICC(2,1) is significantly higher than previous datasets, indicating our annotation process, with clear instructions, a well-defined rubric, and high-quality sentence pairs, promotes more consistent judgments.

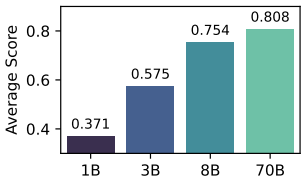

Figure 1: Average human ratings of each model size, normalized to 0-1. We see that larger models consistently receive higher scores.

Moreover, we find that human ratings strongly correlate with model size, as demonstrated in Figure 1. Computing the Spearman rank correlation between human ratings and model size scores yields a correlation of 0.759. This result supports our assumption that model size is a meaningful proxy for simplification quality. Given the limited scale of our human annotated dataset, we aim to further validate the effectiveness of SynthSimpliEval at scale through automated evaluation using LLMs.

4 Auto-Evaluation with LLMs-as-a-Jury on SynthSimpliEval

We now introduce an LLMs-as-a-Jury framework for evaluating text simplification on SynthSimpliEval, given the difficulty and costs of attaining high quality ratings from human raters. In our setup, we follow prior work in which multiple LLMs (as a jury of judges) are prompted to assess the quality of simplifications based on a shared rubric.

Drawing on previous work on text simplification that employs Likert scales for evaluation (Xu et al., 2016; Alva-Manchego et al., 2020), we prompt each LLM in our jury to produce a score from 1 to 5 for each simplification using the same prompt (see Appendix C). We first construct a prompt $P = [I_{pre}; c; I_{mid}; t; I_{post}]$ by concatenation (;), where c is the source sentence and t is the target. I_{mid} and I_{post} are primarily formatting, and I_{pre} is a natural language instruction that elicits some score s from the model. We take these instructions

Type	Sentence
Complex	The integration of advanced payment systems in chemical engineering plants necessitates a robust cybersecurity framework to protect sensitive operational data and financial transactions.
Llama 3 1B	Using advanced payment systems in factories helps keep the important things safe. To keep this safe, strong computer security measures are needed.
Llama 3 3B	To keep information about the plants and money safe, we need a strong system to protect it from hackers and keep track of all the payments.
Llama 3 8B	Using computers in factories that make chemicals requires a strong way to keep the information and money safe from being hacked.
Llama 3 70B	When we use new ways to pay for things in chemical plants, we need to make sure we have strong protection for our computers and money information so it doesn't get stolen or hurt.

Table 4: An example simplification of a synthetic sentence in our dataset by the four Llama 3 models. The simplification prompt can be found in Appendix C.

and query multiple language models (LMs) $J_1, ..., J_n$ to collect a set of scores $S = s_1, ..., s_n$. Using an aggregation function $f(S)$, we compute the final score $q = f(S)$.

We find that few-shot prompt with rationale generation performs best. To enhance reproducibility, we use a lower temperature setting, though not zero, as we observe that small models occasionally produce outputs that are difficult to parse correctly. In these cases, regenerating a response with a different seed addresses the problem. Additionally, we consider a diverse range of instruction-tuned models to appear on our jury, from locally hosted ones to proprietary systems, allowing us to analyze how performance varies across different model architectures and sizes. These models are the instruction-tuned versions of: Gemma 2 27B (Team, 2024), Qwen 2.5 32B (Qwen et al., 2025), Mixtral 8x7B (Jiang et al., 2024), Qwen 2.5 72B, Deepseek V3 (DeepSeek-AI et al., 2025), Claude 3.7 Sonnet (Anthropic, 2025), and GPT-4o (OpenAI et al., 2024). We normalize the collected scores to 0-1, and use the arithmetic mean of all models as our aggregation strategy.

4.1 LLMs-as-a-Jury Ratings In Agreement

Just as it is crucial that human raters produce a clear signal, our evaluation method must also be reliable, i.e., consistent among themselves. We compute $ICC(2,1)$ among the 7 LLM judges, and find that they show strong agreement with each other (Table 2 under LLM Unified). Among the 3 SOTA models — Deepseek V3, Claude 3.7 Sonnet, and GPT-4o) — this is even higher at 0.754. The Spearman rank correlations between individual models can be found in Appendix B.

Simultaneously, many existing text simplification datasets (Alva-Manchego et al., 2021; Maddela et al., 2021) assess simplification quality across three dimensions: fluency, meaning preservation, and simplicity. While this partitioning was useful in earlier evaluation settings, it may no longer be necessary or effective for modern LLMs. In particular, SOTA LLMs rarely produce disfluent outputs, which reduces the informativeness of the fluency score.

To examine the utility of this traditional three-way partition, we compare it to a unified scoring approach in which LLMs are asked to provide a single overall rating of simplification quality. Specifically, we prompt the LLM judges to independently rate fluency, meaning, and simplicity on a 1–5 Likert scale, and compute the inter-rater agreement across models using ICC(2,1). We then compare these values to the ICC(2,1) score obtained when LLM judges give a single unified rating, as in our default setup.

As shown in Table 2, we find that the agreement among LLM judges is consistently higher when using unified scoring when compared to ratings for meaning and simplicity in the three-way partition. This suggests that a unified rating not only simplifies the evaluation process but is also more reliable, offering a more stable signal for simplification quality in the context of LLM-generated outputs.

4.2 LLM-as-a-Jury Ablation Studies

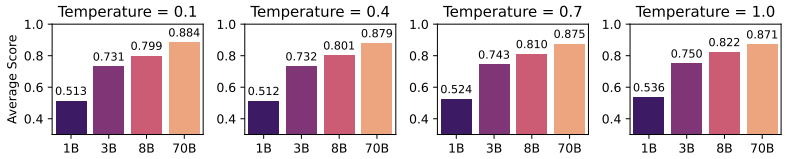

Figure 2: Temperature ablation on SynthSimpliEval. Spearman rank correlations from left to right are 0.626, 0.622, 0.615, and 0.581 respectively.

In this section, we study the impact of various design choices on LLM judgment quality. As shown in §3.2, model size is a proxy for simplification quality. Therefore, we perform ablations to optimize correlation with model size, finding a strong setup for LLM judges that covers model selection, rationale generation, and few-shot prompts.

We perform our ablations with the following base configuration: Our language model judges have a temperature of 0.1 and a consistent prompt found in Appendix C. Our base prompt is few-shot and asks the model to provide a rationale before answering. Using this setup, we test all LLMs in our jury. Also, note that we do not test any Llama models as judges, as they are the simplifiers and may introduce bias through self-evaluation.

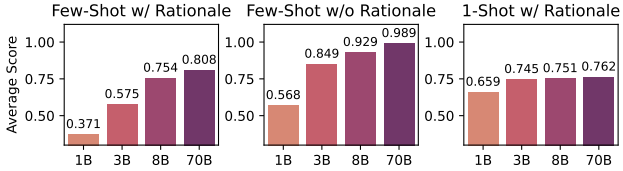

Figure 3: Prompt format ablations on SynthSimpliEval. Spearman rank correlations from left to right are 0.626, 0.335, and 0.648 respectively.

Apart from model selection, our other ablations use Qwen 2.5 72B Instruct as a judge. The 1040 input sentence pairs are as described in §3.1, and we compare the average scores of each of the four simplifier models. We also compute their Spearman rank correlations with the model size score in §3.1.

We find that temperature has minimal impact on judge performance. A slight decrease in correlation is noticed, but this is likely due to randomness from increasing temperature.

On the other hand, we find that few-shot prompts significantly outperforms 1-shot. The latter prompt results in a large percentage of resulting scores being 4 out of 5, likely because all the simplifications are of decent quality. Our few-shot examples may be encouraging the model to be stricter in its judgments, resulting in a flatter distribution of scores. While the correlation with model size is similar, including the rationale depresses the average scores. With rationale, the 70B simplifier model drops from a near-perfect average (without rationale) to 4.52 out of 5. This leaves room for potentially stronger models — such as the 405B variant of Llama 3 — while preserving the model's ability to judge accurately.

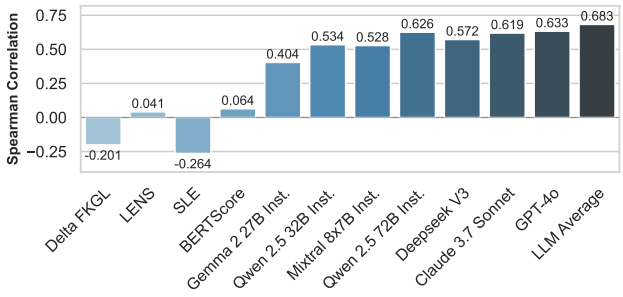

Figure 4: Spearman correlations on SynthSimpliEval between existing metrics, LLMs, and LLM average with model size. See full correlation matrix in Appendix B.

Lastly, we perform an ablation on 7 judge models (§4) of various sizes and report results in Figure 4. We find that larger and closed-source models, on the right, tend to perform better than their smaller counterparts. Notably, the panel's overall judgment — the average score

— performs markedly better than any individual model. While it is difficult to form an exact ranking, as correlation is a proxy, we find that model performance is consistent with general understandings of model output quality.

4.3 Evaluating Existing Metrics and LLMs-as-a-Jury on SynthSimpliEval

First, we use our silver standard benchmark SynthSimpliEval to assess a set of widely used or new automatic evaluation metrics for text simplification: FKGL (Flesch, 1948), LENS (Maddela et al., 2023), BERTScore (Zhang et al., 2019), and SLE (Cripwell et al., 2023). Note that we do not use systems that are significantly dependent on reference sentences, such as SARI, as our synthetic dataset does not include them. Systems that are not totally dependent on references, such as LENS, are included by setting the reference to the simplification itself.

FKGL measures readability using sentence length and syllable count. We calculate Delta FKGL, the difference between simple and expert sentences, to assess relative simplicity. We define it as follows: $\Delta FKGL = FKGL(c) - FKGL(t)$. As a high FKGL represents a difficult sentence, $\Delta FKGL$ is high when the complex sentence is much harder than the simple sentence. LENS uses an encoder transformer, RoBERTa-large, to encode sentences into vectors. A trained feedforward neural network then predicts a single scalar score using these vectors as inputs. Lastly, BERTScore, which is generally used for text similarity, has also been applied to measure text simplicity. BERTScore generally compares word embeddings of the complex, reference, and simple sentences, while BLEU uses a formula that considers n-gram precision and sentence brevity. Similar to FKGL, SLE is an absolute measure. However, it is instead computed by a fine-tuned LLM. We use ΔSLE as defined in Cripwell et al. (2023) to measure a simplification's quality.

We plot the average scores of each of the 4 simplifier models in Figure 5. None of the existing metrics show a strong correlation with the simplifier model's size. On the contrary, FKGL and SLE exhibit an opposite correlation, where larger models receive lower simplification scores than smaller models. While BERTScore and LENS seem to show some correlation with model sizes, average scores are very similar among the four models, making it difficult to differentiate between them. Spearman correlations with model size are also low in Figure 4. This suggests that none of the existing metrics reliably capture the expected trend

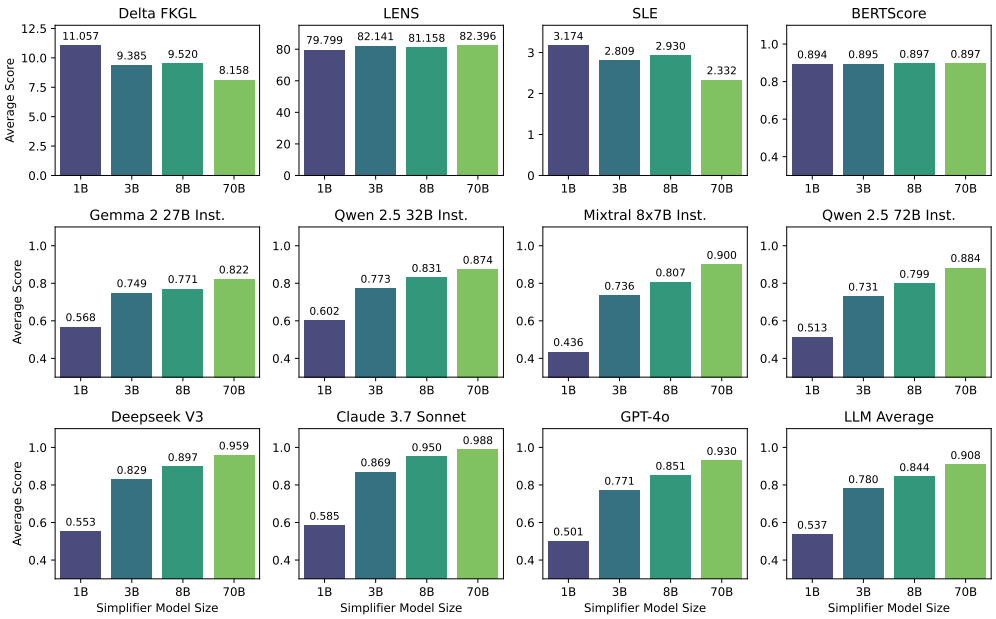

Figure 5: Average scores of each simplifier (Llama 3) model. The top row contains previous metrics, and the bottom two rows contain normalized LLM judge scores (§4).

that previous works suggest and human raters identify. Since existing evaluation metrics fail to reliably distinguish between simplifications generated by models of different sizes, we investigate our alternative approach that uses a panel of LLMs as evaluators.

As shown in Figure 5, both individual LLMs and the aggregated panel consistently assign higher scores to simplifications produced by larger models. This suggests that LLM judges are sensitive to subtle qualitative differences in simplification quality that scale with model size. This indicates that existing simplification metrics — FKGL, LENS, SLE, and BERTScore — struggle to capture the key aspects necessary to evaluate the finer distinctions in LLM-generated simplifications.

5 Can Existing Learnable Metrics Benefit from Synthetic Data?

As mentioned in §2.1, existing datasets generally do not reflect modern simplification systems. Consequently, systems trained on lower quality data — such as LENS (Maddela et al., 2023), SLE (Cripwell et al., 2023), and REFeREE Huang & Kochmar (2024) — may learn information that is less applicable to modern simplification systems. Additionally, the evaluation of multiple LLMs is often expensive, and it would be beneficial to have a smaller model that can perform faster, albeit less accurate, evaluation. We train a small feedforward network with a similar architecture to LENS and SLE to explore these question.

Following existing work, we train a small neural network on the sentence embeddings of the complex and simplified sentences. We collect a separate set of 400 synthetic sentences created in the same manner as SynthSimpliEval (§3.1), each simplified by the 1B and 8B models; this pair of simplifiers produces a diverse range of ratings. While previous work has used models such as RoBERTa, we compute embeddings with the SentenceTransformers library (Reimers & Gurevych, 2019). We use a newer embedding model, `all-mpnet-base-v2`, which is less than half the size of previous work. We compute embeddings of length 768 for the complex and simple sentences, $E_{complex}$ and E_{simple}. The input feature is constructed as $X = [E_{complex}; E_{simple}; E_{complex} - E_{simple}; E_{complex} \odot E_{simple}]$ where $\odot$ is the Hadamard product. Our final network is a single-layer feedforward network with 64 neurons — much smaller than previous work — to predict the score given to the input pair by the Qwen 2.5 72B Instruct judge.

We find that the resulting model has a correlation of 0.22 with the model size score, substantially higher than previous metrics but much lower than the full model. As a small model, this is a strong result which carries the implication that a potential issue facing existing metrics is the quality of their training data.

6 Conclusion

In this work, we address key limitations in existing text simplification benchmarks including low dataset quality and low annotator agreement. We address these limitations by introducing SynthSimpliEval, a synthetic benchmark designed to better reflect the capabilities of modern language models. Our benchmark combines complex inputs from both human-written and model-generated sources, and includes simplifications produced by models of varying sizes. Through a human study with high inter-annotator agreement and correlation with model size, we validate the reliability of our dataset. To scale evaluation and reduce annotation costs, we adopt a panel of LLM judges (LLMs-as-a-Jury), which produces scores that align strongly with model size and show higher consistency than existing metrics. We further demonstrate that learnable metrics can benefit from training on LLMs-as-a-Jury-labeled data, improving their ability to reflect true simplification quality. Overall, our findings suggest a practical and extensible approach for building reliable evaluation resources in tasks where high-quality annotations are limited or unreliable.

References

Suha S Al-Thanyyan and Aqil M Azmi. Automated text simplification: a survey. *ACM Computing Surveys (CSUR)*, 54(2):1–36, 2021.

Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4668–4679, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.424. URL https://aclanthology.org/2020.acl-main.424/.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. The (un)suitability of automatic evaluation metrics for text simplification. *Computational Linguistics*, 47(4):861–889, December 2021. doi: 10.1162/coli_a_00418. URL https://aclanthology.org/2021.cl-4.28/.

Anthropic. Claude 3.7 sonnet and claude code. https://www.anthropic.com/news/claude-3-7-sonnet, February 2025. URL https://www.anthropic.com/news/claude-3-7-sonnet. Accessed: 2025-03-28.

Lora Aroyo and Chris Welty. The three sides of crowdtruth. *Human Computation*, 1(1), 2014.

Lora Aroyo and Chris Welty. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24, 2015.

Bernd Bohnet, Vinh Q Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, et al. Attributed question answering: Evaluation and modeling for attributed large language models. *arXiv preprint arXiv:2212.08037*, 2022.

Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Sheila Castilho. On the same page? comparing inter-annotator agreement in sentence and document level human machine translation evaluation. In Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Yvette Graham, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, and Matteo Negri (eds.), *Proceedings of the Fifth Conference on Machine Translation*, pp. 1150–1159, Online, November 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.wmt-1.137/.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*, 2023.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A survey on evaluation of large language models, 2023. URL https://arxiv.org/abs/2307.03109.

Liam Cripwell, Joël Legrand, and Claire Gardent. Simplicity level estimate (sle): A learned reference-less metric for sentence simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12053–12059, 2023.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report, 2025. URL https://arxiv.org/abs/2412.19437.

Ashwin Devaraj, William Sheffield, Byron Wallace, and Junyi Jessy Li. Evaluating factuality in text simplification. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume*

1: Long Papers), pp. 7331–7345, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.506. URL https://aclanthology.org/2022.acl-long.506/.

Yue Dong, Zichao Li, Mehdi Rezagholizadeh, and Jackie Chi Kit Cheung. EditNTS: An neural programmer-interpreter model for sentence simplification through explicit editing. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3393–3402, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1331. URL https://aclanthology.org/P19-1331/.

Rudolph Flesch. A new readability yardstick. *Journal of applied psychology*, 32(3):221, 1948.

Veronika Hackl, Alexandra Elena Müller, Michael Granitzer, and Maximilian Sailer. Is gpt-4 a reliable rater? evaluating consistency in gpt-4's text ratings. In *Frontiers in Education*, volume 8, pp. 1272229. Frontiers Media SA, 2023.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Pei-Yun Hsueh, Prem Melville, and Vikas Sindhwani. Data quality from crowdsourcing: a study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 workshop on active learning for natural language processing*, pp. 27–35, 2009.

Hui Huang, Yingqi Qu, Xingyuan Bu, Hongli Zhou, Jing Liu, Muyun Yang, Bing Xu, and Tiejun Zhao. An empirical study of llm-as-a-judge for llm evaluation: Fine-tuned judge model is not a general substitute for gpt-4, 2024. URL https://arxiv.org/abs/2403.02839.

Yichen Huang and Ekaterina Kochmar. REFeREE: A REference-FREE model-based metric for text simplification. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 13740–13753, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.1200.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts, 2024. URL https://arxiv.org/abs/2401.04088.

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. Neural crf model for sentence alignment in text simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7943–7960, 2020.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. URL https://arxiv.org/abs/2001.08361.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, et al. Holistic evaluation of language models, 2023. URL https://arxiv.org/abs/2211.09110.

Feng Liu, Tao Xiang, Timothy M Hospedales, Wankou Yang, and Changyin Sun. Inverse visual question answering: A new benchmark and vqa diagnosis tool. *IEEE transactions on pattern analysis and machine intelligence*, 42(2):460–474, 2018.

Mounica Maddela, Fernando Alva-Manchego, and Wei Xu. Controllable text simplification with explicit paraphrasing. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3536–3553, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.277. URL https://aclanthology.org/2021.naacl-main.277.

Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. LENS: A learnable evaluation metric for text simplification. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16383–16408, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.905. URL https://aclanthology.org/2023.acl-long.905.

Ian R McKenzie, Alexander Lyzhov, Michael Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Aaron Kirtland, Alexis Ross, Alisa Liu, et al. Inverse scaling: When bigger isn't better. *arXiv preprint arXiv:2306.09479*, 2023.

Shashi Narayan and Claire Gardent. Hybrid simplification using deep semantics and machine translation. In *The 52nd annual meeting of the association for computational linguistics*, pp. 435–445, 2014.

Stefanie Nowak and Stefan Rüger. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the international conference on Multimedia information retrieval*, pp. 557–566, 2010.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, et al. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

Arjun Panickssery, Samuel R Bowman, and Shi Feng. Llm evaluators recognize and favor their own generations. *arXiv preprint arXiv:2404.13076*, 2024.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. Mauve: Measuring the gap between neural text and human text using divergence frontiers, 2021. URL https://arxiv.org/abs/2102.01454.

Maja Popović, Sheila Castilho, Rudali Huidrom, and Anya Belz. Reproducing a manual evaluation of the simplicity of text simplification system outputs. In Samira Shaikh, Thiago Ferreira, and Amanda Stent (eds.), *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*, pp. 80–85, Waterville, Maine, USA and virtual meeting, July 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.inlg-genchal.12/.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL http://arxiv.org/abs/1908.10084.

Alex Reinhart, Ben Markey, Michael Laudenbach, Kachatad Pantusen, Ronald Yurko, Gordon Weinberg, and David West Brown. Do llms write like humans? variation in grammatical and rhetorical styles. *Proceedings of the National Academy of Sciences*, 122(8):e2422455122, 2025.

Rion Snow, Brendan O'connor, Dan Jurafsky, and Andrew Y Ng. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pp. 254–263, 2008.

Elior Sulem, Omri Abend, and Ari Rappoport. BLEU is not suitable for the evaluation of text simplification. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 738–744, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1081. URL https://aclanthology.org/D18-1081.

Teerapaun Tanprasert and David Kauchak. Flesch-kincaid is not a text simplification evaluation metric. In Antoine Bosselut, Esin Durmus, Varun Prashant Gangal, Sebastian Gehrmann, Yacine Jernite, Laura Perez-Beltrachini, Samira Shaikh, and Wei Xu (eds.), *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pp. 1–14, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.gem-1.1. URL https://aclanthology.org/2021.gem-1.1.

Gemma Team. Gemma. 2024. doi: 10.34740/KAGGLE/M/3301. URL https://www.kaggle.com/m/3301.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. Replacing judges with juries: Evaluating llm generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*, 2024.

Junlin Wang, WANG Jue, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities. In *The Thirteenth International Conference on Learning Representations*, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Xuanxin Wu and Yuki Arase. An in-depth evaluation of gpt-4 in sentence simplification with error-based human assessment. *arXiv preprint arXiv:2403.04963*, 2024.

Sander Wubben, Antal van den Bosch, and Emiel Krahmer. Sentence simplification by monolingual machine translation. In Haizhou Li, Chin-Yew Lin, Miles Osborne, Gary Geunbae Lee, and Jong C. Park (eds.), *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1015–1024, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL https://aclanthology.org/P12-1107/.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3: 283–297, 2015. doi: 10.1162/tacl_a_00139. URL https://aclanthology.org/Q15-1021/.

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

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

Xingxing Zhang and Mirella Lapata. Sentence simplification with deep reinforcement learning. *arXiv preprint arXiv:1703.10931*, 2017.

A Related Work

Existing text simplification metrics broadly fall into two categories: static and learnable metrics. Traditional metrics were deterministic, and often depended on word or n-gram occurrence. Examples of this include SARI (Xu et al., 2016) and BLEU (Papineni et al., 2002). Both SARI and BLEU consider n-gram similarity, and SARI further considers the importance of added and removed n-grams through the use of references. Even earlier approaches included FKGL (Flesch, 1948), which is still commonly used. This computes text simplicity using a formula containing average syllables per word and words per sentence. However, these metrics are not designed for the text simplification task, having been adapted from other fields such as machine translation; recent work has shown that this has limitations, such as negative correlations with simplicity on certain datasets (Sulem et al., 2018; Tanprasert & Kauchak, 2021).

More recently, work has been done on using LMs to measure text simplicity. While initially designed for semantic similarity, BERTScore (Zhang et al., 2019) has been used to measure some aspects of text simplification. More recent work, such as LENS (Maddela et al., 2023), REFeREE (Huang & Kochmar, 2024), and SLE (Cripwell et al., 2023), have trained smaller models (such as RoBERTa) to predict scores. While they perform relatively well, they are also limited by the need to collect datasets with human ratings. SLE circumvents this by using a combination of Newsela data Xu et al. (2015) — already labeled by difficulty — and interpolation with FKGL, but this dataset is also constrained by the generalization of the former and performance of the latter.

Our work builds on language model inference techniques. We base our reasoning on chain-of-thought (Wei et al., 2022), adapted to a classification task, and use few-shot learning (Brown, 2020); in particular, one-shot learning greatly improves performance. Lastly, we use models as evaluators, which have previously shown performance competitive with, and in some cases superior to, human judgment (Bohnet et al., 2022). Additionally, pre-trained models are able to generalize better than their fine-tuned counterparts (Huang et al., 2024). However, one main drawback is that these models tend to prefer their own outputs (Panickssery et al., 2024). To counteract this, we use juries as proposed by Verga et al. (2024) to improve performance while decreasing hardware requirements and costs. We additionally take steps to ensure that models judge neither their own outputs nor the outputs of other models in their family.

B Full Correlation Matrix

To compute the full correlation matrix, each of the existing metrics and LLM judges score the 1040 data pairs in SynthSimpliEval, and we compute the Spearman rank correlation. We additionally add the average LLM score and the model size score. The various LLMs have high correlations with each other as well as with the model size score. On the other hand, existing metrics do not correlate well with either.

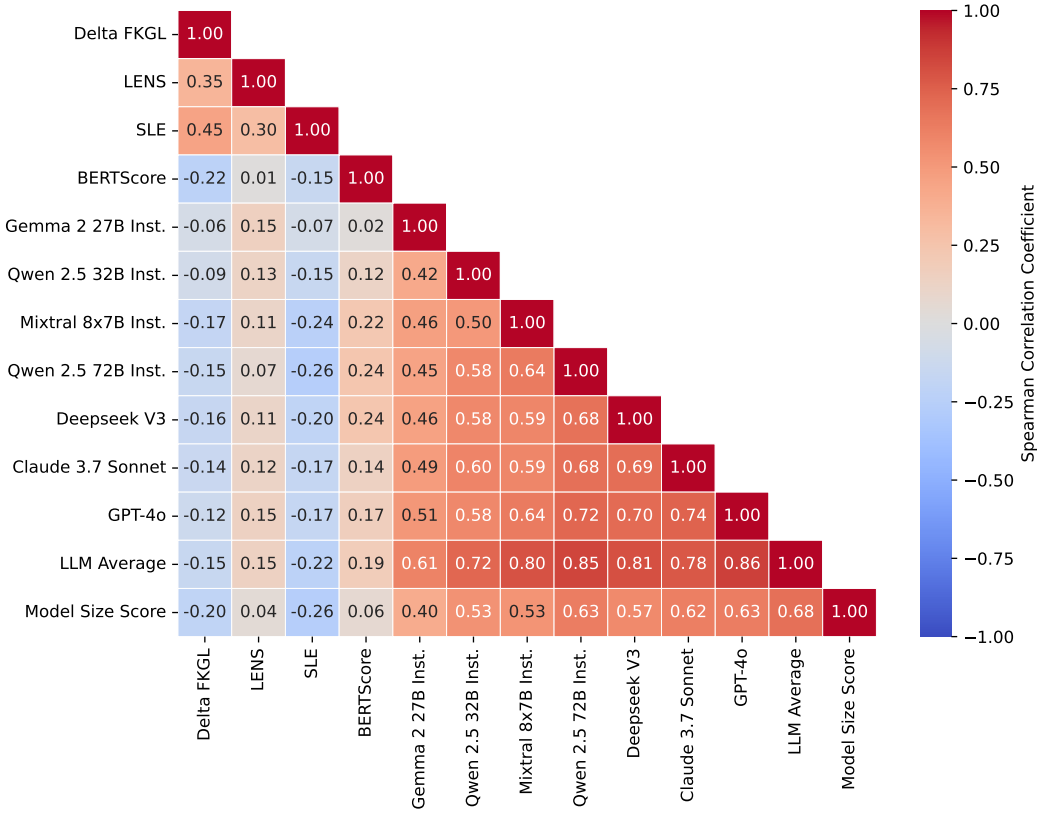

Figure 6: The full correlation matrix between existing metrics, LLM judges, their average, and the model size score from §4.3.

C Prompt Details

We use a chat format for all of our inference tasks. Some prompts have a system role message containing instructions. For models that don't support a system role, we simply prepend it to the first message. The prompts provided are in the ChatML format, but are replaced automatically with whichever format the model defines for vLLM. We additionally have prompts for our various ablations; due to length, however, these can be found in our GitHub repository.

C.1 Synthetic Data Generation

We begin our data generation by asking for complex sentences about a subject in a domain of knowledge (Appendix D). This is a 1-shot prompt, with no system role.

```
<|im_start|>user
Please provide a technically difficult sentence about physical education. The
```

```
sentence should be concise but specific; instead of overcomplicating, try to come
up with something that would be found in a technical report or paper. If you wish,
you may consider the following subject, which may or may not be related: 'school'.
<|im_end|>

<|im_start|>assistant
The PE department's allocation of resources must ensure adequate supervision of
students engaging in high-impact aerobic activities.
<|im_end|>

<|im_start|>user
Please provide a technically difficult sentence about {subject}. The
sentence should be concise but specific; instead of overcomplicating, try to come
up with something that would be found in a technical report or paper. If you wish,
you may consider the following subject, which may or may not be related:
'{seed_noun}'.
<|im_end|>
```

We then ask the various Llama models to simplify the resulting sentences with the following prompt. We use four instruction-tuned models from the Llama 3 family: Llama 3.2 1B Instruct, Llama 3.2 3B Instruct, Llama 3.1 8B Instruct, and Llama 3.3 70B Instruct. We found that a 2-shot prompt produced more reliable outputs, particularly on smaller models.

```
<|im_start|>system
As an expert educator, your task is to simplify the provided sentences into a form
that elementary school children would find easier to understand. In your
simplification, try to make the sentence easier to understand while retaining the
original meaning. Replace technical terms with easier words, particularly those
that are more common. If needed, break the original sentence into multiple
sentences to simplify sentence structure - it is perfectly fine to extend the
length of the output if it makes it easier to understand. Your output will be
judged on 3 aspects: fluency (is it grammatically correct; does it flow well?),
meaning (does it preserve the original meaning and nuance?), and most importantly
simplicity (is it easier to read and understand?). Output only the simplified
sentence(s) on a single line.
<|im_end|>

<|im_start|>user
Please simplify the following sentence:\n
The cumulative effect of instructors' accrued sick leave and retirement benefits
must be reconciled with the department's annual expenditure on personnel-related
overheads.
<|im_end|>

<|im_start|>assistant
The money used for teacher's sick time and retirement pay has to match up with the
money the department spends on all its workers every year.
<|im_end|>

<|im_start|>user
Please simplify the following sentence:\n
The prolonged exposure to screens can lead to a significant alteration in the
dopaminergic reward system, potentially affecting cognitive and behavioral
outcomes in adolescents.
<|im_end|>

<|im_start|>assistant
Looking at screens for a long time can change how your brain feels happy and
rewards you. This might affect how young people think and behave.
```

```
<|im_end|>

<|im_start|>user
Please simplify the following sentence:\n
{sentence}
<|im_end|>
```

C.2 LLM Judge Prompt

We provide our final LLM judge prompt. This is a few-shot prompt with a system message.

```
<|im_start|>system
You are an expert linguist and educator. Your task is to judge how well a system
has simplified a sentence. To do this, you will be given a base sentence, which
will be more complex, and a new version of that sentence. In your judgement,
consider text simplicity (how hard is it for children to read?), semantic
similarity (does it mean the same thing, and convey the same nuance?), and fluency
(the simplification should be grammatically correct). First, concisely consider
all three aspects. Note that a sentence can do well in some aspects and badly in
others - use your judgement to decide which is more important for that sentence.
If the sentences are identical, your answer should be around a [3] - but also
consider whether its possible to simplify it further. At the end, provide an
integer score from [1] to [5], where [5] represents a high quality simplification,
in square brackets. Provided below is a general rubric:\n
[5]: The sentence is fully simplified, entirely fluent, and preserves the core
meaning of the original sentence.\n
[4]: The sentence is somewhat simpler, mostly fluent, and the meaning is close to
the original sentence.\n
[3]: The sentence is simpler, somewhat fluent, and the meaning is similar to the
original sentence.\n
[2]: The sentence is equivalently simple, still has some fluency, but the meaning
is lost.\n
[1]: The sentence is completely unreadable.
<|im_end|>

<|im_start|>user
Base Sentence: It will then dislodge itself and sink back to the river bed in
order to digest its food and wait for its next meal.\n
New Sentence: It will then move away from the river bed and sink back to the
bottom to digest its food.
<|im_end|>

<|im_start|>assistant
The sentence reads completely fluently. The original meaning of the sentence -
something unattaching and sinking to the bottom - is altered, as the subject
"moves away" from the river bed in the simple sentence as opposed to moving
towards it. It is somewhat simpler than the original, such as by using "move away"
instead of "dislodge". Answer: [3]
<|im_end|>

<|im_start|>user
Base Sentence: As a result, although many mosques will not enforce violations,
both men and women when attending a mosque must adhere to these guidelines.\n
New Sentence: Due to this, a lot of mosques don't enforce these rules but both men
and women should follow them.
<|im_end|>

<|im_start|>assistant
Although the new sentence has an error in fluency (missing a comma after "these
```

```
rules"), it has the same meaning as the original. The sentence also uses simpler
vocabulary (such as by using "follow" instead of "adhere") and sentence
structure, notwithstanding the grammatical error. Answer: [4]
<|im_end|>

<|im_start|>user
Base Sentence: A Wikipedia gadget is a JavaScript and/or a CSS snippet that can be
enabled simply by checking an option in your Wikipedia preferences.\n
New Sentence: A a gadget javascript a is and / checking wikipedia an sinppet that
can be enabled simply by or css option in your wikipedia preferences.
<|im_end>

<|im_start|>assistant
The new sentence is completely unreadable. Answer: [1]
<|im_end|>

<|im_start|>user
Base Sentence: Despite what many consider obvious, that which we pursue with
greatest fervor often eludes our grasp not for lack of effort, but precisely
because our desperate clutching drives away the very thing we seek.\n
New Sentence: Despite what many consider obvious, that which we pursue with
greatest fervor often eludes our grasp not for lack of effort, but precisely
because our desperate clutching drives away the very thing we seek.
<|im_end|>

<|im_start|>assistant
The new sentence is equivalent in simplicity, fluency, and meaning to the
original. However, there are many ways to improve the sentence's readability.
Answer: [3]
<|im_end|>

<|im_start|>user
Base Sentence: The dog ran across the green field to catch the ball.\n
New Sentence: The dog ran across the green field to catch the ball.
<|im_end|>

<|im_start|>assistant
The new sentence is equivalent in simplicity, fluency, and meaning to the
original. However, it would be difficult to simplify this sentence further, as it
is already very straightforward. Answer: [4]
<|im_end|>

<|im_start|>user
Base Sentence: {base_sentence}\n
New Sentence: {new_sentence}
<|im_end|>
```

D Synthetic Dataset Generation Lists

D.1 Domains of Knowledge

We select 20 domains of knowledge to ensure our datasets cover a wide range of potential sentences. These are:

Domains of Knowledge			
architecture	chemical engineering	physics	chemistry
art	nursing	music	biology
mathematics	philosophy	theater	anthropology
english	education	accounting	history
computer science	political science	economics	psychology

D.2 Concept Nouns

Our concept nouns were collected by filtering a list of common English nouns by removing words with common verb or adjective forms. The full list of 739 filtered concept nouns is as follows.

Concept Nouns				
people	history	art	world	information
map	family	government	system	computer
meat	year	music	person	method
data	food	theory	law	bird
literature	problem	software	knowledge	ability
economics	internet	television	science	library
fact	product	idea	temperature	investment
society	activity	story	industry	thing
oven	community	definition	safety	quality
development	language	management	player	variety
video	country	exam	movie	organization
equipment	physics	analysis	policy	series
direction	strategy	technology	army	camera
freedom	environment	child	month	truth
university	writing	article	department	difference
goal	audience	growth	income	marriage
user	combination	failure	medicine	philosophy
teacher	communication	chemistry	disease	energy
nation	road	soup	location	success
apartment	education	painting	politics	decision
event	property	student	wood	competition
distribution	entertainment	office	population	president
unit	category	cigarette	context	introduction
opportunity	performance	driver	flight	length
magazine	newspaper	relationship	cell	dealer
finding	lake	member	phone	scene
association	concept	customer	discussion	housing
inflation	insurance	woman	effort	expression
importance	opinion	payment	reality	responsibility
situation	skill	wealth	application	city
county	depth	estate	foundation	grandmother
perspective	photo	recipe	studio	topic
collection	depression	imagination	resource	agency
college	connection	criticism	debt	description
patience	secretary	solution	administration	director
personality	psychology	recommendation	selection	alcohol
complaint	contract	highway	loss	membership
possession	preparation	steak	union	agreement
cancer	currency	employment	engineering	interaction
mixture	region	republic	tradition	virus
actor	classroom	delivery	device	difficulty
drama	election	engine	football	guidance
hotel	owner	protection	suggestion	variation
anxiety	atmosphere	awareness	bath	bread
candidate	comparison	confusion	construction	elevator
emotion	employee	employer	guest	leadership
mall	manager	operation	recording	sample
transportation	charity	cousin	disaster	editor
efficiency	excitement	guitar	homework	leader
outcome	presentation	promotion	refrigerator	resolution

revenue	session	singer	tennis	basket
bonus	cabinet	childhood	church	clothes
dinner	drawing	initiative	judgment	lab
measurement	mud	poetry	police	possibility
procedure	queen	relation	restaurant	satisfaction
sector	signature	significance	song	tooth
town	vehicle	volume	wife	accident
airport	arrival	baseball	chapter	committee
conversation	database	enthusiasm	explanation	farmer
gate	girl	hall	historian	hospital
injury	instruction	manufacturer	meal	perception
pie	poem	proposal	reception	replacement
revolution	river	son	speech	village
winner	worker	writer	assistance	buyer
chest	chocolate	conclusion	contribution	cookie
courage	desk	drawer	establishment	examination
garbage	grocery	improvement	independence	insect
inspection	inspector	king	ladder	penalty
piano	potato	profession	professor	quantity
requirement	salad	sister	supermarket	weakness
wedding	ambition	analyst	apple	assignment
assistant	bathroom	bedroom	celebration	championship
cheek	client	consequence	departure	diamond
dirt	fortune	friendship	gene	girlfriend
hat	lady	negotiation	obligation	passenger
pizza	platform	poet	pollution	recognition
reputation	shirt	speaker	stranger	surgery
tale	trainer	uncle	youth	film
water	money	example	business	study
game	field	fish	experience	job
book	economy	body	market	state
radio	company	card	list	group
force	key	training	school	research
service	web	boss	sport	house
page	soil	oil	picture	garden
site	exercise	image	case	coast
action	boat	result	section	building
mouse	cash	class	store	tax
space	rule	model	source	earth
program	chicken	purpose	question	rock
salt	birth	car	dog	object
scale	sun	war	bank	craft
bus	eye	fire	box	frame
step	cycle	metal	room	screen
structure	ball	discipline	gift	machine
tool	career	culture	pot	sign
table	task	egg	ice	network
star	challenge	brush	plant	wing
brain	button	foot	wall	distance
pair	savings	staff	sugar	target
animal	author	budget	file	ground
lesson	officer	sky	stage	stick
title	bowl	bridge	campaign	character
club	evidence	fan	letter	novel
park	quarter	baby	dish	fruit
glass	muscle	strength	vegetable	chart
gear	kitchen	land	log	mother
relative	street	tree	bench	commission
path	project	sea	ticket	confidence
daughter	doctor	dot	duty	essay
father	milk	pipe	seat	stable
storm	substance	team	bat	beach
chain	consideration	cream	crew	gold
interview	kid	mission	shop	suit
window	agent	band	block	bone

calendar	cap	coat	contest	court
cup	district	door	finger	garage
hole	hook	layer	lecture	meeting
nose	rice	telephone	airline	bag
battle	bed	cake	designer	dimension
dress	emergency	extension	farm	horror
horse	husband	mountain	nail	noise
occasion	package	patient	phrase	sand
sentence	stomach	string	tourist	towel
vacation	wheel	wine	arm	associate
border	branch	brother	coach	document
expert	floor	god	iron	judge
knife	landscape	league	parent	pin
pool	pound	salary	shelter	shoe
tank	bell	bike	boy	brick
chair	closet	clue	collar	conference
devil	glove	jacket	monitor	mortgage
nurse	peak	plane	reward	sandwich
yard	bicycle	bottle	cable	candle
clerk	cloud	concert	counter	flower
grandfather	lawyer	mirror	pension	plate
ruin	ship	skirt	snow	specialist
trash	anger	award	boot	bug
camp	candy	carpet	cat	champion
clock	cow	engineer	entrance	grass
incident	island	jury	leg	lip
motor	nerve	passage	pen	priest
prize	resident	resort	ring	roof
rope	scheme	script	sock	station
toe	tower	truck	witness	human
individual	guard	watch	official	press
spring	objective	chemical	dump	conflict
mobile	train	bear	representative	

E Existing Dataset Samples

We provide randomly selected additional samples of complex-simple sentence pairs from both previous works and SynthSimpliEval.

Source	Complex Sentence	Simplified Sentence
Simplicity-DA	These works he produced and published himself, whilst his much larger woodcuts were mostly commissioned work.	These works he made and published himself and his much larger woodcuts were written work.
Simplicity-DA	The SAT Reasoning Test (formerly Scholastic Aptitude Test and Scholastic Assessment Test) is a standardized test for college admissions in the United States.	The SAT Reasoning Test (used to be called Scholastic Aptitude Test and Scholastic Assessment Test) is a test for college admissions in the United States.
Newsela-Likert	president barack obama understands that if he were to proclaim a goal of definitively eliminating isis in the short term, he would fail.	president barack obama knows that he can not promise to destroy the islamic state quickly. he would fail.
Newsela-Likert	he could not move or talk and he looked like he was sleeping.	he was hurt badly and could not move or even open his eyes.
SimpEval2022	Two sisters, Leah and Chantrelle, and their acquaintance Hosanna catch a steamboat from Saint Ann Parish in Jamaica to the United Kingdom, arriving in London's Notting Hill before moving to the Midlands.	Two sisters, Leah and Chantrelle, as well as their friend Hosanna catch a boat from Jamaica to London. Then, they will move to the midlands.
SimpEval2022	Drone footage released by the Islamic State showed bombs being dropped on an ammunitions facility located in Deir ez-Zor, Syria, an area of contested control between the Islamic State and the Syrian government at the time.	The Islamic State and the Syrian government were fighting to control an area in Syria called Deir ez-Zor. During that time, the Islamic State released videos that showed bombs being dropped on a weapons storage facility in that area.
SynthSimpliEval	The historical development of elevator technology in urban high-rise buildings significantly impacted the architectural design and social stratification of cities in the early 20th century.	In the early 1900s, buildings with high ceilings became common in cities. This made the people who lived in high buildings feel like they were better than those in lower buildings, and it changed the way people's homes were designed.
SynthSimpliEval	The significance of accurate variance analysis in financial reporting is paramount, as it directly impacts the reliability of financial statements and the decision-making processes of stakeholders.	The accuracy of financial reports is very important. Without it, people don't trust the information and can't make smart decisions about the company's business.

Table 6: Additional sentence pairs