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

Joseph Liu, Yoonsoo Nam, Xinyue Cui, Swabha Swayamdipta
Department of Computer Science
University of Southern California
{jliu7350, yoonsoon, xinyuecu, swabhas}@usc.edu

Abstract

Static benchmarks, or fixed datasets created once and applied repeatedly, are still the default choice for evaluating language models, despite two major challenges. First, static benchmarks rarely reflect evolving model capabilities, often containing outdated examples that are too easy, disfluent, or incoherent. Second, existing human ratings associated with these benchmarks often contain a high degree of disagreement, resulting in inconsistent ratings; existing metrics must nevertheless correlate with these ratings. This hurts evaluation reliability and might break expected trends (e.g., more powerful models being assigned higher scores). We address these challenges, using the task of text simplification as a case study, through two contributions. First, we introduce SynthSimpliEval, a static 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. Overall, through our case study, we show that a reliable evaluation requires higher quality test data in a static benchmark, which could be obtained through careful collection of synthetic data and LLMs-as-a-Jury ratings.

1 Introduction

Static benchmarks remain the predominant approach for evaluating LLMs; they require only one-time data collection, minimal evaluation infrastructure, and enable highly standardized comparisons across studies. However, they face two limitations. First, data often does not qualitatively match the performance of modern models due to the latter's continuous improvement, a problem made worse by potentially low-quality samples resulting from the extensive use of crowdsourcing in initial data collection. Second, the human ratings included in the benchmarks often exhibit low inter-rater agreement, creating noisy and inconsistent ground truth. Dynamic benchmarks [Kiela et al., 2021], such as WebArena for web-based tasks [Zhou et al., 2024] and LiveCodeBench for code generation [Jain et al., 2024], require substantial computational resources, infrastructure, and maintenance, making them impractical.

Text simplification, an essential task for improving information accessibility, exemplifies these challenges. The gold standard for evaluation is human judgment [Devaraj et al., 2022, Maddela et al., 2023], but existing human ratings show low inter-annotator agreement [Wu and Arase, 2024, Popović et al., 2022] and rely on data that is low quality and unrepresentative of modern LLM outputs.

To address these challenges, we propose SynthSimpliEval, a synthetic benchmark featuring complex sentences simplified by LLMs of varying sizes. Our approach leverages the observation that within

Accepted to NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle.

model families, larger models consistently outperform smaller ones across NLP tasks [Hestness et al., 2017, Kaplan et al., 2020, Hoffmann et al., 2022, Liang et al., 2023, McKenzie et al., 2023]. Through a pilot study, we validate that human ratings on our benchmark exhibit high inter-annotator agreement and correlate strongly with model size, confirming larger models produce higher-quality simplifications; model size creates clearer quality gradations than scraped and formatted internet text, making human ratings less ambiguous.

We scale our evaluation beyond costly human annotation by adopting an LLMs-as-a-Jury approach [Verga et al., 2024, Chan et al., 2023, Wang et al., 2024], where multiple models evaluate simplification quality. Our systematic study shows that few-shot prompting with rationale generation yields scores that correlate better with model size (an established indicator of model power than existing metrics.

We show that existing text simplification metrics [Flesch, 1948, Maddela et al., 2023, Zhang et al., 2019, Cripwell et al., 2023], when applied to SynthSimpliEval, fail to reflect expected quality trends. However, LLMs-as-a-Jury consistently assigns higher scores to more powerful models' outputs, aligning with quality trends across model scales and providing stronger evidence of its validity as an evaluation method than existing metrics.

2 Existing Text Simplification Benchmarks and Ratings

Evaluating text simplification metrics requires datasets with complex-simple sentence pairs and human ratings [Xu et al., 2015, 2016, Maddela et al., 2023]. Given a source sentence c and target simplification t, evaluation methods compute a score $q(c, t)$ that should correlate with human ratings. However, existing benchmarks face two critical challenges that undermine reliable evaluation.

Coherence and Difficulty Gaps in Existing Datasets. Existing benchmarks, such as Simplicity-DA [Alva-Manchego et al., 2021], Newsela-Likert [Maddela et al., 2021], and SimpEval2022 [Maddela et al., 2023], contain simplifications that poorly reflect modern LLM capabilities. Many include grammatical errors, disfluencies, and overly simplistic outputs from outdated systems (PBMT-R, Hybrid, SBMT-SARI for Simplicity-DA; EditNTS, LSTM for Newsela-Likert). These failure modes are rare in current LLMs [Reinhart et al., 2025], creating a mismatch between benchmark errors and contemporary system outputs. Additionally, existing datasets primarily use sentences from Wikipedia and news articles that often lack technical complexity. They fail to test LLM capabilities on domain-specific content requiring specialized vocabulary simplification or complex syntactic restructuring, where nuance preservation and accuracy are critical [Devaraj et al., 2022].

Annotator Disagreement. Human ratings in existing datasets show poor inter-annotator agreement. In Newsela, the Intraclass Correlation Coefficient (ICC[2,1]) is low across three separate dimensions—fluency (0.153), meaning (0.213), and simplicity (0.054). A unified rating approach, as used in SimpEval2022, shows similar issues, with an ICC of just 0.228. This disagreement stems from the subjective nature of "simplicity," which varies by educational background and domain knowledge [Aroyo and Welty, 2015], and from inconsistent annotation guidelines that often leave raters to define simplicity themselves [Alva-Manchego et al., 2021].

These fundamental problems with existing benchmarks, including poor data quality and unreliable human annotations, motivate our synthetic benchmark with controlled quality variations.

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, SynthSimpliEval, constructed with Qwen 2.5 72B Instruct and Llama 3 (§3.1). We assess the quality of our benchmark on the basis of two criteria: agreement among human annotators, demonstrating reliability; and performance of models of different sizes(§3.2).

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.

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 1: 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.

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 Appendix G 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 1, our dataset is comparable or slightly larger than prior works. We then use model size as a direct proxy for simplification quality, motivated by prior findings that large LLMs generally produce higher-quality outputs (§1), by selecting simplification models of increasing size: Four Llama 3 models (1B to 70B). 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 40 randomly selected complex sentences from SynthSimpliEval (20 each from SimpEval2022 and our synthetic dataset), paired with simplifications from four models (totaling 160 pairs). Three human annotators rate simplification quality on a 5-point Likert scale using consistent instructions (Appendix C.2). Their inter-annotator agreement measured in ICC(2,1) is significantly higher (0.627) than previous datasets, indicating our annotation process — with clear instructions and rubrics, and high-quality sentence pairs — promotes more consistent judgments.

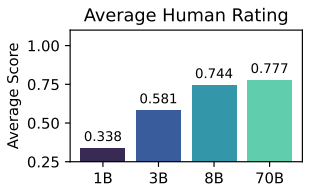

Figure 1: Average human ratings of each model size, normalized to 0-1. We see that larger models consistently receive higher scores.

Moreover, we find that human ratings strongly correlate with model size (Figure 1); Spearman rank correlation between human ratings and model size scores is 0.707. 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 further validate the effectiveness of SynthSimpliEval at scale through automated evaluation using LLMs.

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

Given the difficulty and costs of obtaining reliable human ratings, we adopt an LLMs-as-a-Jury framework [Verga et al., 2024] for evaluating text simplification on SynthSimpliEval. Multiple LLMs serve as judges, rating simplifications on a 1-5 Likert scale using a consistent rubric (see Appendix C for prompts).

Our jury consists of seven instruction-tuned models: 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]. Each model rates simplifications using few-shot prompting with rationale generation. We aggregate scores by taking the arithmetic mean across all models.

4.1 LLMs-as-a-Jury Shows Strong Agreement

Our evaluation method produces reliable signals. LLM judges show strong inter-rater agreement with ICC(2,1) = 0.641 for unified ratings on SynthSimpliEval, substantially higher than human agreement on prior benchmarks. Agreement is even stronger among SOTA models (see Appendix B).

We also compare unified scoring against the traditional three-dimensional approach (fluency, meaning, simplicity). When LLM judges rate Newsela using both approaches, we find comparable ICC scores: 0.672 (fluency), 0.650 (meaning), 0.640 (simplicity) for the three-dimensional approach versus 0.657 for unified scoring. While agreement is similar, unified scoring simplifies the evaluation process and is more appropriate for modern LLMs that rarely produce disfluent outputs.

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

Building on our human validation that larger models produce better simplifications, We evaluate LLMs-as-a-Jury as well as existing metrics, including FKGL [Flesch, 1948], LENS [Maddela et al., 2023], BERTScore [Zhang et al., 2019], and SLE [Cripwell et al., 2023], on SynthSimpliEval by measuring correlation with model size as a proxy for quality.

First, to verify that LLMs produce meaningful scores, we compute correlation between human and LLM scores on the human-rated subset (§3.2). The averaged human score and the averaged LLM score have a Spearman rank correlation of 0.672.

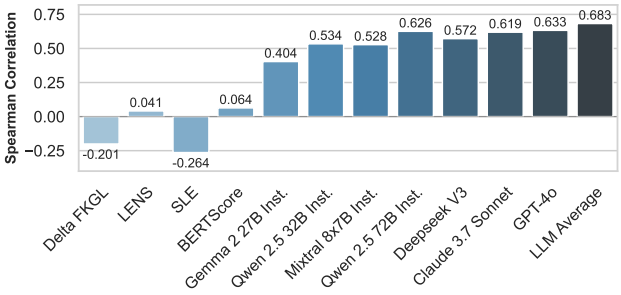

Figure 2: Spearman correlations on SynthSimpliEval between existing metrics, LLMs, and LLM average with model size. See full correlation matrix in Appendix B.

Figure 2 shows that existing metrics fail to capture expected quality trends. FKGL and SLE show negative correlations, incorrectly rating larger models as producing worse simplifications. While BERTScore and LENS show weak positive correlations, their scores barely differentiate between model sizes. In contrast, both individual LLMs and the aggregated panel consistently assign higher scores to larger models, successfully capturing quality differences that align with human judgments.

These results demonstrate that LLMs-as-a-Jury provides a more reliable evaluation than existing metrics for modern text simplification systems. Detailed ablation studies on temperature, prompting strategies, and judge selection are provided in Appendix E.

5 Conclusion

We address key limitations in existing text simplification benchmarks, including low dataset quality and low annotator agreement, through two main contributions. First, we introduce SynthSimpliEval, a synthetic benchmark that better reflects modern LLM capabilities. Through human evaluation, we show high inter-annotator agreement (ICC = 0.627) and validate that larger models produce better simplifications. Second, we show that LLMs-as-a-Jury provides reliable evaluation that aligns with both model size and human ratings, outperforming existing automatic metrics. Synthetic generation is extensible to variants of text simplification, such as document-level simplification, and potentially to other tasks. 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

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. 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.

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.

Liam Cripwell, Joël Legrand, and Claire Gardent. Simplicity level estimate (sle): A learned referenceless metric for sentence simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 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, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 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/.

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

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.

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, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13740–13753, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.1200.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*, 2024.

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.

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.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Talat, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. Dynabench: Rethinking benchmarking in nlp. *ArXiv*, abs/2104.14337, 2021. URL https://api.semanticscholar.org/CorpusID:233444226.

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.

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, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 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, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 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.

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*, pages 311–318, 2002.

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, editors, *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*, pages 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.

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.

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, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 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, editors, *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 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.

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.

Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, and Xianyi Cheng. Tianyue ou, yonatan bisk, daniel fried, uri alon, and graham neubig. webarena: A realistic web environment for building autonomous agents. In *The Twelfth International Conference on Learning Representations*, volume 2, page 8, 2024.

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 and 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 and 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.

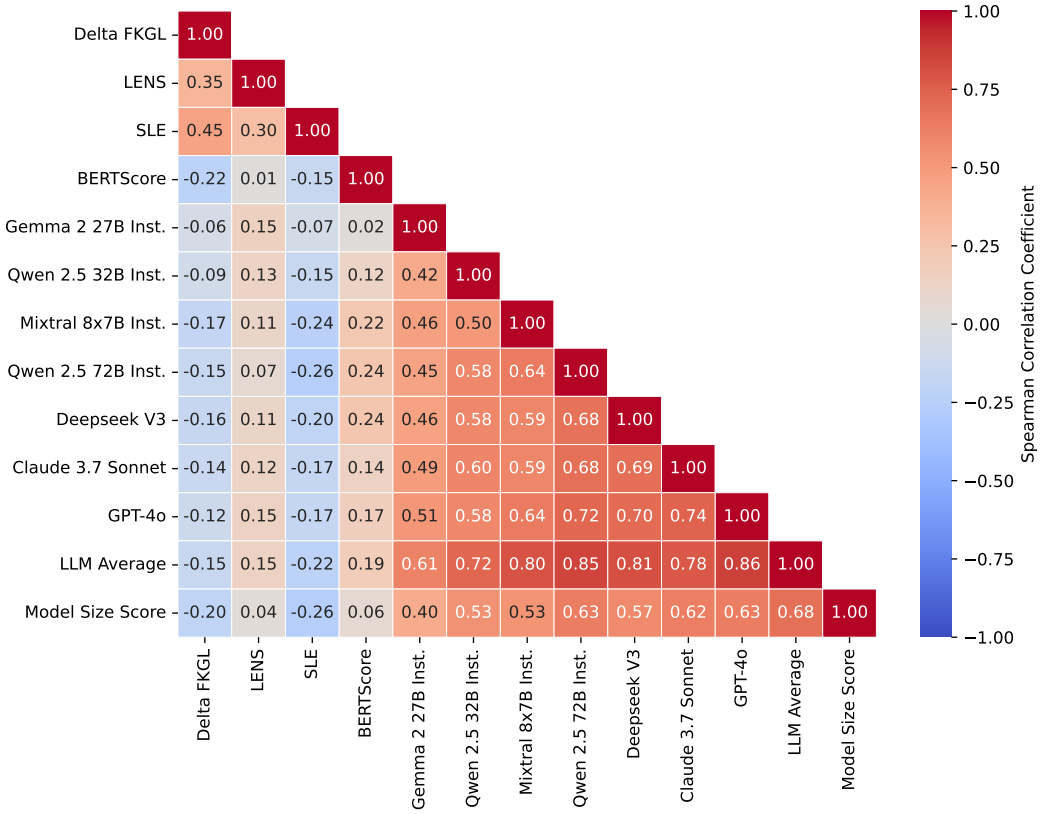

Figure 3: The full correlation matrix between existing metrics, LLM judges, their average, and the model size score from §4.2.

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 LLM-as-a-Jury Ablation Studies

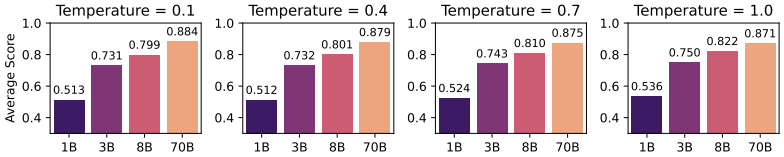

Figure 4: Temperature ablation on SynthSimpliEval. Spearman rank correlations from left to right are 0.626, 0.622, 0.615, and 0.581 respectively.

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. 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.

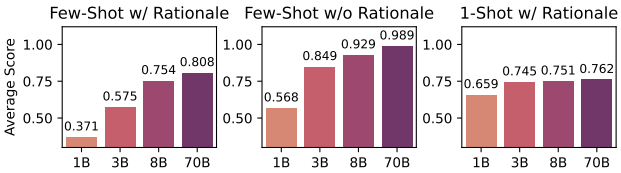

Figure 5: Prompt format ablations on SynthSimpliEval. Spearman rank correlations from left to right are 0.626, 0.648, and 0.335 respectively.

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 stronger future models while preserving the current models' ability to judge accurately.

Lastly, we perform an ablation on 7 judge models (§4) of various sizes and report results in Figure 2. 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 for quality, we find that model performance is consistent with general understandings of model output quality.

F 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 3: Additional sentence pairs

G Example Synthetic Sentences

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.

Table 4: Synthetic complex sentence samples from SynthSimpliEval.