EasyJon at TSAR 2025 Shared Task: Evaluation of Automated Text Simplification with LLM-as-a-Judge

Paul-Gerhard Barbu[1,3], Adrianna Lipska-Dieck[2], Lena Lindner[1],

[1]Rosenheim University of Applied Sciences,
[2]German Center for Neurodegenerative Diseases (DZNE) Dresden
[3]EasyJon,

Correspondence: paulgbarbu@gmail.com

Abstract

This paper presents an approach to automated text simplification for CEFR A2 and B1 levels using large language models and prompt engineering. We evaluate seven models across three prompting strategies: short, descriptive, and descriptive with examples. A two-round evaluation system using LLM-as-a-Judge and traditional metrics for text simplification determines optimal model-prompt combinations for final submissions. Results demonstrate that descriptive prompts consistently outperform other strategies across all models, achieving 46-65% of first-place rankings. Qwen3 shows superior performance for A2-level simplification, while B1-level results are more balanced across models. The LLM-as-a-Judge evaluation method shows strong alignment with traditional metrics while providing enhanced explainability.

1 Introduction

A crucial component of inclusion lies in language accessibility, where barriers to text readability and comprehension remain significant (Saggion, 2017; Rennes, 2022). Enhancing technological inclusion requires improving text readability, which shapes how easily diverse audiences can process, understand, and engage with written material (Vajjala and Meurers, 2014; Schriver, 1990; Saggion, 2017). Text simplification, whether achieved through manual strategies (e.g., via trained editors) or automated approaches (e.g., large language models), is a key method for enhancing readability and comprehension by adjusting content to the target audience (Saggion, 2017). Within this context, automated, readability-controlled text simplification using large language models (LLMs) and refined prompt engineering represents an innovative approach. It offers a scalable, cost-efficient, and adaptive solution that can be tailored to specific readability levels and target audiences (Chen, 2025; Kew et al., 2023).

The central challenge in automated, readability-controlled text simplification lies in balancing reduced linguistic complexity with the preservation of meaning and fluency (Agrawal and Carpuat, 2024). Achieving this often requires supervision mechanisms and nuanced adjustments rather than straightforward simplification (Agrawal and Carpuat, 2024). LLMs, particularly when guided by adequate prompt engineering strategies, offer a promising avenue by adapting text to specific proficiency levels while retaining semantic accuracy and naturalness (Agrawal and Carpuat, 2024; Paulson and Hernandez, 2025; Chen et al., 2025; Barayan et al., 2025). The evaluation of the quality of the output poses similar challenges and requires similar finesse (Vajjala, 2022).

We employ a LLM-as-a-Judge (Zheng et al., 2023) approach to evaluate the simplifications alongside the metrics provided by the TSAR Team (Alva-Manchego et al., 2025).

Our approach builds upon the simplification tool EasyJon by (Barbu, 2024) which employs a LLM-as-a-Judge approach to find the best performing model and prompt technique for simplifying text into German plain language *Leichte Sprache*.

The aim of this study was to create the most target-level appropriate text simplification with the use of LLMs and non-complex prompt engineering as well as compare two different evaluation approaches of LLM-based automated text simplifications.

2 Methodology

The task was completed in two distinct phases: the *simplification* process, which involved multiple prompts and LLMs, and the subsequent *evaluation* of the appropriate simplifications for the final submission. For the simplification, we tested three prompts and seven models. For the evaluation, we employed an LLM-as-a-Judge system (Zheng et al., 2023). The dataset was provided by the TSAR 2025 shared task team (Alva-Manchego et al., 2025).

2.1 LLM Models

We simplified the texts with seven models of which five were open-weight models. **Qwen3 235B A22B** (Yang et al., 2025), **Llama 3.3 70B Instruct**[1] – which is based on its predecessor Llama 3.1 (Grattafiori et al., 2024), **DeepSeek R1 Distill Llama 70B** (DeepSeek-AI et al., 2025), **Gemma 3 27B** (Team et al., 2025), **GPT OSS 120B** (OpenAI et al., 2025) and two closed source models **Claude Sonnet 4** (by Anthropic[2]) and **Mistral Medium 3.1** (by Mistral[3]). For inference we used OpenRouter[4].

2.2 The simplification prompts

Our simplification approach built upon the work of Barayan et al. (2025), adopting their prompting strategies: short, descriptive, and descriptive + example. We employed more recent LLMs and adapted the prompt content to address limitations in the original study. Finally, we implemented a novel evaluation framework to provide more comprehensive assessment.

We developed six prompts following a consistent structure (See A.1): three for A2-level simplification and three for B1-level simplification. The complete prompt instructions were detailed in the appendices (for A2 in Appendix A.2 and for B1 in A.3). All models were configured with a conservative temperature setting of 0.3 to ensure consistent output.

2.3 Evaluating text simplifications with an LLM-as-a-Judge system and traditional metrics

To determine the best simplification for each text, we employed an LLM-as-a-Judge system with a two-round tournament. In the first round, the three prompts competed within each model. The simplification generated by each prompt was then assessed for quality using LLM-as-a-Judge and placed to compete in round two. In the second round, the winning simplifications competed across models. The winners from round two formed our final submission dataset, resulting in a heterogeneous combination of simplifications from different models and prompts.

2.4 The judgement prompt

Claude Sonnet-4 was chosen to take on the role of the judge based on the strong performance of its predecessor (Sonnet 3.5) demonstrated by (Barbu, 2024) for assessing readability consistency on the CLEAR Dataset (Crossley et al., 2022). The judgment prompt was manually adjusted specifically for this evaluation task. (Detailed in A.4)

The LLM-as-a-Judge was employed identically in both rounds. For each round the judge was used to evaluate the strengths and weaknesses of each simplification and rank them from best to worst. The evaluation prompt consisted of the task, the original text and the simplifications. (Example in Appendix A.5)

To mitigate potential bias that the LLM-as-a-Judge might have exhibited toward specific models, we implemented pseudo-anonymization. Model names were abbreviated to only their first two letters before being presented to the judge (e.g., "Model LL" for Llama 3.3 80B Instruct).

2.5 Scoring strategy

To enable quantitative analysis, we converted the LLM-as-a-Judge's rankings into numerical scores using a Borda count approach. The best-ranked simplification received 6 points, the second-best received 5 points, and so on down to 0 points for the worst-ranked simplification. This conversion from ordinal rankings to numerical scores allowed us to calculate means and standard deviations across multiple evaluation rounds.

3 Results

3.1 Results of round one

The assessment of the prompt strategy involved aggregating the first-place rankings assigned by the LLM-as-a-Judge across all 100 texts at each target level (A2 and B1). For each of the 200 original texts to be simplified, three different prompting strategies were applied, and the LLM-as-a-Judge evaluated which strategy produced the best simplification.

Figures 1 and 2 present the frequency with which each prompting strategy achieved first place

[1] https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct
[2] https://docs.anthropic.com/en/docs/about-claude/models/overview
[3] https://docs.mistral.ai/getting-started/models/models_overview/
[4] https://openrouter.ai/

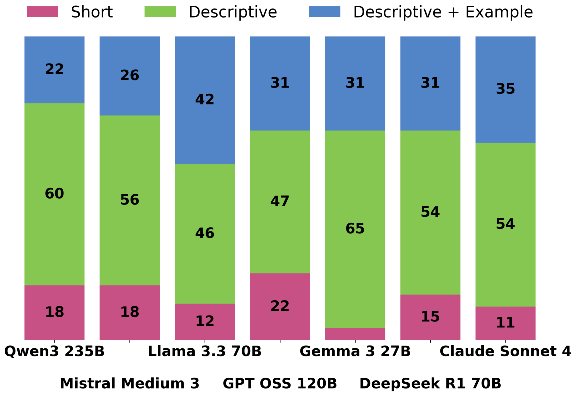

Figure 1: First-place rankings of prompting strategies across seven models for A2

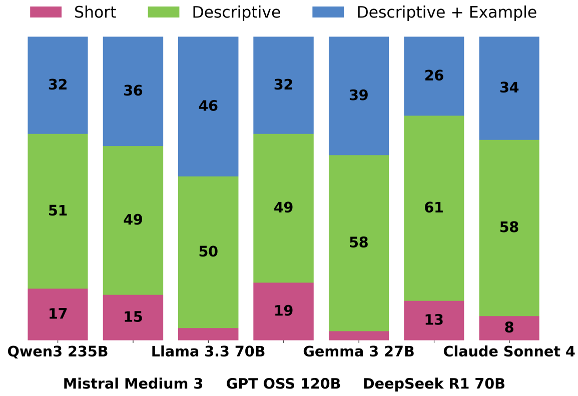

Figure 2: First-place rankings of prompting strategies across seven models for B1

across the 100 texts at each simplification level, evaluated across seven different models. The results demonstrate that the descriptive prompt achieved the highest number of first-place rankings, consistently outperforming other prompts across all tested models for both A2 and B1 simplification levels. Specifically, the descriptive prompt secured between 46 and 65 first-place rankings out of 100 texts depending on the model, representing the most successful simplification strategy. The descriptive with example emerged as the second-best performing prompt (22 to 46 first-place rankings out of 100), while the short prompt consistently yielded the fewest top rankings across all models. This pattern remained remarkably consistent across diverse model architectures, from smaller models like Gemma 3 27B to larger systems such as Claude Sonnet 4, suggesting that detailed task descriptions provided more effective guidance for text simplification regardless of model capacity or design. However, the inclusion of just one example may have proven more restrictive than beneficial.

3.2 Results of round two

While in round one the task was evaluating the prompts for a model, round two examined the overall performance patterns across models by calculating the mean and standard deviation of the Borda score for each model. The results revealed distinct performance patterns between the A2 and B1 simplification levels as presented in Table 1. For A2 simplifications, Qwen3 demonstrated superior performance with a mean Borda score of 5.07, establishing a substantial gap ahead of the second-ranked model, Mistral Medium 3, which achieved a mean score of 3.39. This considerable margin suggested Qwen3's particular effectiveness in A2-level text simplification tasks. In contrast, the B1 simplification results exhibited greater balance across models. It is possible that B1 simplifications present less complexity, potentially allowing each model to achieve competitive performance.

3.3 Qualitative Analysis of simplifications

This section presents two example simplifications. We selected the shortest A2 text along with two of its simplifications produced by Qwen 3 and Gemma 3. The two models that advanced to the second round of competition. To demonstrate the spectrum of performance, we chose the highest-ranked and lowest-ranked simplifications according to the LLM-as-a-Judge evaluation for this exact comparison.

Original: Now NASA is working towards logging some of the smaller asteroids, those measuring 140 metres wide or more. Of the 25,000 estimated asteroids of this size, so far about 8,000 have been logged, leaving 17,000 unaccounted for. Considering that a 19-metre asteroid that exploded above the city of Chelyabinsk in Russia in 2013 injured 1,200 people, these middle-sized asteroids would be a serious danger if they enter Earth's orbit.

Qwen 3: NASA is now trying to record smaller asteroids, especially those 140 meters wide or bigger. Scientists think there are about 25,000 of these asteroids, but so far they have only found around 8,000. That means about 17,000 are still missing. In 2013, a much smaller asteroid—only 19 meters wide—exploded above Chelyabinsk, Russia, and hurt 1,200 people. So, if one of these larger asteroids came near Earth, it could be very dangerous.

Model	A2	B1	**A2 & B1**
Qwen3 235B	**5.07±1.62**	3.37±1.80	**4.22±1.71**
Mistral Medium 3	3.39±1.50	**3.58±2.01**	3.49±1.75
Llama 3.3 70B	3.38±1.47	3.19±1.56	3.29±1.51
GPT OSS 120B	2.90±1.76	3.11±2.14	3.00±1.95
Claude Sonnet 4	2.75±1.97	3.17±2.13	2.96±2.05
DeepSeek R1 70B	2.71±1.73	2.54±2.04	2.62±1.89
Gemma 3 27B	0.80±1.27	2.04±1.91	1.42±1.59

Table 1: Mean and standard deviation of the Borda scores based on the ranking by the LLM-as-a-Judge system. Higher Borda scores indicate a better text simplification performance. Bold values indicate the best performance for the category (sorted by A2 & B2)

Gemma 3: Now NASA is trying to find and record smaller asteroids. Asteroids are big rocks in space. They are looking for asteroids that are 140 metres wide or bigger. Scientists think there are about 25,000 asteroids of this size. So far, they have found about 8,000. That means there are still 17,000 asteroids they haven't found yet. In 2013, a smaller asteroid – about 19 metres wide – exploded over a city in Russia called Chelyabinsk. More than 1,200 people were hurt by the explosion. So, these middle-sized asteroids could be very dangerous if they come close to Earth.

LLM-as-a-Judge evaluation for Qwen 3: Strengths: Maintains appropriate A2 vocabulary while preserving technical accuracy. Uses "record" as a good simplification of "logging." Includes all key information including the specific location "Chelyabinsk, Russia." Clear logical flow between sentences.

Weaknesses: Uses "missing" which might be slightly confusing since the asteroids aren't lost, just untracked.

LLM-as-a-Judge evaluation for Gemma 3: Strengths: Provides clear definition "Asteroids are big rocks in space" which is very helpful for A2 learners. Uses simple vocabulary and sentence structures consistently. Includes specific location details.

Weaknesses: The definition placement creates awkward sentence flow. Some sentences are choppy. Uses "More than 1,200" instead of the exact figure.

3.4 Comparing LLM-as-a-Judge with traditional metrics

The LLM-as-a-Judge system demonstrated the strongest alignment with the MeaningBERT-to-Reference (mb-ref) metric when examining model rankings, as can be seen in table 2. This becomes

Model	RMSE	mb-org	**mb-ref**
Reference	*0.612*	*0.808*	-
Qwen3 235B	0.825	0.845	0.842
Mistral Medium 3	0.725	0.819	0.839
DeepSeek R1 70B	0.857	0.845	0.835
GPT OSS 120B	0.803	0.838	0.830
Claude Sonnet 4	0.735	0.812	0.822
Llama 3.3 70B	0.667	0.802	0.821
Gemma 3 27B	0.725	0.789	0.811

Table 2: Models evaluated by the metrics provided by (Alva-Manchego et al., 2025) ordered by MeaningBERT-Reference.

evident when models are ordered from best to worst performance according to their mb-ref scores. The ranking produced by the LLM-as-a-Judge system closely mirrors the ordering derived from mb-ref values. However, the observed alignment may not have represented a significant relationship, given that similarity scores fall within a relatively narrow range. The observed span from 0.811 to 0.842 represented a modest variation of only 0.031 points, suggesting that while the ordering appeared consistent, the practical differences between model performances is minimal.

4 Conclusion

In this study, we aimed to create the most appropriate simplifications for CEFR A2 and B1 target levels, as well as compare two evaluation approaches for LLM-based automated text simplification. We conclude that the descriptive prompting strategy is optimal for obtaining CEFR target level text simplifications. We also conclude that the evaluation strategy of using an LLM-as-a-Judge proves to be equally proficient as the traditional evaluation metrics for text simplification.Qwen3 demonstrated superior performance for A2-level tasks, whereas Mistral Medium emerged as the top-performing model for B1 simplifications. Notably, the performance distribution across models was more balanced for B1-level tasks compared to A2-level tasks. The primary advantage of employing an LLM-as-a-Judge approach lies in its explainability. Unlike traditional metrics, an LLM can identify, penalize, and provide detailed explanations for problematic simplifications. For instance, it can detect awkward phrasing, identify idioms requiring advanced comprehension despite simple vocabulary, and flag complex terms that are immediately clarified by an explanation. This approach offers significant utility for dataset creation, fine-tuning smaller models, and diagnosing model performance issues in an interpretable manner. The explainable feedback enables researchers to pinpoint specific simplification shortcomings, whether they stem from inappropriate phrasing, word choice, sentence structure, or other linguistic factors. Such granular insights are crucial for iterative model improvement and understanding the nuanced challenges in text simplification. An example evaluation demonstrating this approach can be found in Section A.5 in the Appendix. The complete dataset is available on GitHub[5].

Examining additional automated readability and simplification assessment strategies could be a valuable area for research. Comparing our prompt refinement strategy to the findings of further studies would also provide useful insights. Furthermore, assessing text readability not only for target proficiency levels but also for specific audiences, such as individuals with cognitive impairments, neurodivergent individuals, or non-native speakers, may yield important insights.

Limitations

Models may be susceptible to the specific prompts provided. Tailoring a prompt for each model, rather than employing a one-prompt-fits-all-models approach, might yield overall better results. Furthermore, we employed the Borda count method to establish a ranking system for comparative model evaluation. While this approach provided a straightforward solution, alternative ranking methods may offer different perspectives on model and prompt performance.

[5] https://github.com/PaulGBarbu/TSAR2025

Lay Summary

Our research explores how artificial intelligence (AI) can simplify complex texts to make them easier to understand for people at different reading levels. We focused on creating simplified versions of texts suitable for learners at two specific English proficiency levels: A2 (elementary) and B1 (intermediate).

How did we do it? We tested seven different AI's, giving each one three types of instructions: short and simple instructions, detailed instructions, or detailed instructions with examples. Each AI model then simplified 200 texts (100 for each reading level). To find the best simplifications, we used an innovative two-stage evaluation process. First, we compared the three different instruction types within each AI model. Then, we compared the winning simplifications across all seven models. For evaluation, we used another AI system called "LLM-as-a-Judge" (like having an AI referee) alongside traditional measurement methods.

What did we find? The results showed that detailed instructions worked best - consistently producing 46-65% of the top-ranked simplifications across all models. Short instructions performed poorly, and surprisingly, adding examples to the instructions didn't help as much as expected and may have even limited the AI's flexibility. For elementary-level (A2) simplifications, one model called Qwen3 performed significantly better than others. For intermediate-level (B1) texts, the results were more balanced across different models, suggesting that B1 simplification may be an easier task.

Why does this matter? This research has important implications for accessibility and inclusion. Many people struggle with complex text, including non-native speakers, people with cognitive difficulties, and those with lower literacy levels. AI-powered text simplification offers a cost-effective, scalable way to make information more accessible to everyone. The study also demonstrates that the AI evaluation method (LLM-as-a-Judge) can be as effective as traditional metrics while providing better explanations. Unlike traditional measurements, AI judges can explain why a simplification is good or bad - for example, identifying awkward phrasing or unnecessarily complex vocabulary. This feedback is valuable for improving AI systems and creating better training data.

Acknowledgments

Our thanks goes to Mirian Anschütz for the enriching feedback on the simplification made by EasyJon during my (Paul) Master Thesis, which greatly motivated me to pursue further research into this field. We are especially grateful to Marcel Tilly and Gerd Beneken for providing the opportunity to develop this project at TH Rosenheim. Your unwavering support and encouraging feedback have been invaluable throughout this journey. We would like to express our sincere gratitude to Johann Jordan for the for graciously allowing us to adopt his nickname for our tool name. We also wish to thank Andrea Halbritter for her thoughtful article about our tool, which provided constructive feedback and inspired us to continue refining our work. Finally, we are grateful to the reviewers whose insightful feedback and comments have enabled us to strengthen and improve this work.

References

Sweta Agrawal and Marine Carpuat. 2024. Do text simplification systems preserve meaning? a human evaluation via reading comprehension. *Transactions of the Association for Computational Linguistics*, 12:432–448.

Fernando Alva-Manchego, Regina Stodden, Joseph Marvin Imperial, Abdullah Barayan, Kai North, and Harish Tayyar Madabushi. 2025. Findings of the TSAR 2025 shared task on readability-controlled text simplification. In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025)*, Suzhou, China. Association for Computational Linguistics.

Abdullah Barayan, Jose Camacho-Collados, and Fernando Alva-Manchego. 2025. Analysing zero-shot readability-controlled sentence simplification. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6762–6781, Abu Dhabi, UAE. Association for Computational Linguistics.

Paul-Gerhard Barbu. 2024. Entwicklung einer Anwendung zum Übersetzten von Texten in leichter/einfacher Sprache mithilfe von Large Language Models (LLMs). Master's thesis, Rosenheim Technical University of Applied Sciences. Advised and supervised by Prof. Dr. Gerd Beneken and Prof. Dr. Marcel Tilly.

Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. 2025. Unleashing the potential of prompt engineering for large language models. *Patterns*, 6(6):101260.

Jie Chen. 2025. Ai-driven text simplification : Improving web accessibility for cognitive challenges.

Scott Crossley, Aron Heintz, Joon Suh Choi, Jordan Batchelor, Mehrnoush Karimi, and Agnes Malatinszky. 2022. A large-scaled corpus for assessing text readability. *Behavior Research Methods*, 55(2):491–507.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Tannon Kew, Alison Chi, Laura Vásquez-Rodríguez, Sweta Agrawal, Dennis Aumiller, Fernando Alva-Manchego, and Matthew Shardlow. 2023. Bless: Benchmarking large language models on sentence simplification. *Preprint*, arXiv:2310.15773.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, and 108 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Dave Paulson and Lucas Hernandez. 2025. Evaluating language models for simplifying health literacy materials. *Preprints*.

Evelina Rennes. 2022. *Automatic adaptation of Swedish text for increased inclusion*. Linköping University Electronic Press, Linköping.

Horacio Saggion. 2017. *Automatic Text Simplification*. Springer International Publishing.

Karen Schriver. 1990. Evaluating text quality: The continuum from text-focused to reader-focused methods. *Professional Communication, IEEE Transactions on*, 32:238 – 255.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Sowmya Vajjala. 2022. Trends, limitations and open challenges in automatic readability assessment research. *Preprint*, arXiv:2105.00973.

Sowmya Vajjala and Detmar Meurers. 2014. Readability assessment for text simplification: From analyzing documents to identifying sentential simplifications. *International Journal of Applied Linguistics, Special Issue on Current Research in Readability and Text Simplification*, 165.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

A Appendix

A.1 Prompt structure for A2 and B1 simplification

```
[Instruction]
<Input>
[Original Text]
</Input>
Answer just with the simplification
and nothing else. Keep the original
tone.
```

A.2 Instructions for A2

These are the instructions for simplifying to CEFR Level A2.

Short: Simplify the following input to A2 CEFR Level.

Descriptive: You are a text simplification AI. Your task is to simplify the following input to A2 CEFR level. Use only common, everyday words that are appropriate for the context. Choose words that native speakers would naturally use. Explain essential terms if those can't be simplified and maintain the content as in the original.

Descriptive with one example: You are a text simplification AI. Your task is to simplify the following input to A2 CEFR level. Use only common, everyday words that are appropriate for the context. Choose words that native speakers would naturally use. Explain essential terms if those can't be simplified and maintain the content as in the original. Here is a example:

Original
The Villa des Arts (Villa of the Arts) is a complex of artists' studios in the 18th arrondissement of Paris, built at the end of the 19th century. The entrance is located at 15 Rue Hégésippe-Moreau. The nearest metro stations are La Fourche on line 13 and Place de Clichy on line 2. The apartments and studios, grouped around several courtyards with gardens, were listed as a Monument historique in 1994. A2
The Villa des Arts is a place where artists work in Paris. It is in the 18th district of Paris. People built it at the end of the 19th century. The entrance is at Rue Hégésippe-Moreau number 15. If you want to go there by metro, you can use station La Fourche (line 13) or station Place de Clichy (line 2) because they are close. The building has many apartments and studios where artists work. There are also gardens around the buildings. In 1994, the government made this building a historical monument. This means it is on a special list to protect it.

A.3 Instructions for B1

The instruction for simplifying to CEFR Level B1.

Short: Simplify the following input to B1 CEFR Level.

Descriptive: You are a text simplification tool for B1 level. Your task is to simplify the following input to B1 CEFR level. Use clear, straightforward language that intermediate learners can understand. Choose commonly used words and expressions that sound natural. When you must use more complex terms that are important to the topic, provide brief explanations. Keep all the main ideas and information from the original text.

Descriptive with one example: You are a text simplification tool for B1 level. Your task is to simplify the following input to B1 CEFR level. Use clear, straightforward language that intermediate learners can understand. Choose commonly used words and expressions that sound natural. When you must use more complex terms that are important to the topic, provide brief explanations. Keep all the main ideas and information from the original text. Here is a example:

Original
The benefits of remote work are undeniable. Employees save considerable time and money by eliminating their daily commute, while

companies can reduce overhead costs associated with maintaining large office spaces. Many workers report higher productivity levels when working from their home environment, free from office distractions and unnecessary meetings. The flexibility to manage personal responsibilities alongside professional duties has proven particularly valuable for parents and caregivers. B1

The good things about working from home are clear. Workers save a lot of time and money because they don't need to travel to the office every day. Companies also spend less money because they don't need big offices. Many people say they work better at home because there are fewer things that stop them from focusing, and they don't have so many meetings they don't need. Being able to take care of family things while also doing their job has been very helpful, especially for parents and people who look after family members.

A.4 Prompt for LLM-as-a-Judge

The following text was translated to the CEFR [A2|B1] Level by different models. Give feedback to all of, stay factual and dont use words like 'excellent' or similar words, just state the strengths and weaknesses of the different tranlations. Then rank them from best to worst translation. Format the ranking as ["model_i, "model_j","model_k"] at the end of your answer.

A.5 Example Evaluation in Round Two

Text_ID: 24-a

Prompt: `The following text was translated to the CEFR A2 Level by different models. Give feedback to all of, stay factual and dont use words like 'excellent' or similar words, just state the strengths and weaknesses of the different tranlations. Then rank them from best to worst translation. Format the ranking as ["model_i, "model_j","model_k"] at the end of your answer.`

Original

Chloe has two pet rats, Zena and Roxy. Chloe knows they're not typical pets and she has lost count of the times she's patiently tried to explain to people that rats aren't dirty and they make fantastic pets if you look after them correctly. But she understands why people react badly. She didn't like rats either until her friend Martin showed her his pet rats and she saw how affectionate and intelligent they were. Last year, when one of Martin's rats had babies, Chloe decided she'd like to have two of them.

llama-3.3-70b-instruct

Chloe has two pet rats, Zena and Roxy. She knows that most people don't keep rats as pets, and she's often explained to them that rats are clean and nice pets if you take good care of them. But she understands why people don't like them. Chloe didn't like rats either, until her friend Martin showed her his rats. She saw that they were friendly and smart. When Martin's rats had babies, Chloe wanted two of them as pets.

qwen3-235b-a22b-2507

Chloe has two pet rats named Zena and Roxy. She knows rats are not common pets, and she's lost count of how many times she's had to explain to people that rats aren't dirty. She tells them that rats can be great pets if you take good care of them. But she also understands why some people don't like rats — she didn't like them either, at first. Then her friend Martin showed her his pet rats. She saw how smart and loving they were, and that changed her mind. Last year, when one of Martin's rats had babies, Chloe decided she wanted two of them for herself.

Model	**rmse**	**mb-org**	**mb-ref**
llama-3.3-70b-instruct_descriptive_w_example	**0.5874**	0.8028	0.8171
gemma-3-27b-it_descriptive_w_example	0.6205	0.7963	0.8219
mistral-medium-3_descriptive_w_example	0.6595	0.7955	0.8322
mistral-medium-3_short	0.7246	0.8260	0.8403
mistral-medium-3_winner	0.7246	0.8191	0.8388
qwen3-235b-a22b-2507_descriptive_w_example	0.7517	0.8337	0.8389
llama-3.3-70b-instruct_winner	0.6671	0.8015	0.8208
mistral-medium-3_descriptiv	0.7517	0.8118	0.8353
qwen3-235b-a22b-2507_winner	0.8246	0.8452	**0.8422**
qwen3-235b-a22b-2507_descriptiv	0.8246	0.8428	0.8420
llama-3.3-70b-instruct_descriptiv	0.6782	0.7886	0.8128
claude-sonnet-4_short	0.6856	0.7875	0.8150
claude-sonnet-4_winner	0.7348	0.8123	0.8220
gpt-oss-120b_short	0.800	0.8399	0.8315
deepseek-r1-distill-llama-70b_descriptive_w_example	0.8185	0.8499	0.8336
deepseek-r1-distill-llama-70b_short	0.8155	0.8433	0.8338
claude-sonnet-4_descriptive_w_example	0.7071	0.8013	0.8138
qwen3-235b-a22b-2507_short	0.8602	**0.8602**	0.8414
submitted dataset	0.8216	0.8383	0.8357
gpt-oss-120b_winner	0.8031	0.8378	0.8304
claude-sonnet-4_descriptiv	0.7649	0.8182	0.8206
deepseek-r1-distill-llama-70b_descriptiv	0.8367	0.8420	0.8325
gemma-3-27b-it_short	0.7036	0.7754	0.8109
deepseek-r1-distill-llama-70b_winner	0.8573	0.8447	0.8347
gemma-3-27b-it_winner	0.7246	0.7894	0.8107
gpt-oss-120b_descriptiv	0.8246	0.8265	0.8212
llama-3.3-70b-instruct_short	0.7416	0.7840	0.7990
gemma-3-27b-it_descriptiv	0.7616	0.7818	0.8048
gpt-oss-120b_descriptive_w_example	0.8093	0.8042	0.8082

Table 3: These are the metrics calculated on all the simplifications done by all models and prompts. The underlined entries show the dataset of the best simplifications on a model basis, evaluated by the LLM-as-a-Judge.

gpt-oss-120b

Chloe has two pet rats. Their names are Zena and Roxy. She knows rats are not usual pets. She has tried many times to tell people that rats are not dirty and can be great pets if you care for them. She understands why some people do not like rats. Chloe did not like rats at first. Then her friend Martin showed her his rats. She saw that they are friendly and smart. Last year, Martin's rat had babies. After that, Chloe decided she wanted two rats of her own.

deepseek-r1-distill-llama-70b

Chloe has two pet rats named Zena and Roxy. She knows they are not usual pets. She can't remember how many times she has tried to explain to people that rats are not dirty and are great pets if you take good care of them. She knows why people don't like them. She didn't like rats either until her friend Martin showed her his pet rats. She saw how loving and smart they were. Last year, when one of Martin's rats had a litter, Chloe decided to take two.

mistral-medium-3

Chloe has two pet rats named Zena and Roxy. She knows rats aren't common pets, and she's often tried to tell people that rats aren't dirty. They can be great pets if you take good care of them. But she gets why some people don't like them. She didn't like rats either until her friend Martin showed her his pet rats. She saw how loving and smart they were. Last year, when one of Martin's rats had babies, Chloe decided she wanted two of them.

claude-sonnet-4

Chloe has two pet rats, Zena and Roxy. Chloe knows they are not normal pets. Many times, people say rats are dirty. Chloe tells them that rats are not dirty and they are very good pets if you take care of them well. But she understands why people don't like rats. Chloe didn't like rats before too. Then her friend Martin showed her his pet rats. She saw that rats are loving and smart animals. Last year, one of Martin's rats had babies. Chloe wanted to have two of them.

gemma-3-27b-it

Chloe has two pet rats called Zena and Roxy. She knows rats aren't usual pets, and she's explained to many people that rats are clean and can be great pets if you care for them well. But she understands why some people don't like them. She didn't like rats at first either. Then her friend Martin showed her his pet rats, and she saw they were loving and clever. Last year, Martin's rat had babies, and Chloe decided she wanted two of them.

LLM-as-a-Judge Assessment

I'll analyze each translation for A2 level appropriateness, accuracy, and clarity.

Model LL: Uses simple vocabulary and sentence structures appropriate for A2. Maintains accuracy while simplifying complex phrases like "lost count" to "often explained." The flow is natural and all key information is preserved.

Model QW: Retains more complex elements like "lost count" which may be challenging for A2 learners. Uses longer sentences and more sophisticated vocabulary ("common pets," "changed her mind"). Accurate but potentially above A2 level in complexity.

Model GP: Uses very simple sentence structures with clear, short sentences. Vocabulary is basic and appropriate for A2. However, some awkward phrasing appears ("After that, Chloe decided") and the flow feels choppy due to overly simplified sentence structure.

Model DE: Contains vocabulary that may be too advanced for A2 ("litter" instead of "babies," "can't remember"). The sentence "She can't remember how many times" is more complex than needed. Otherwise maintains good accuracy.

Model CL: Uses simple vocabulary and sentence structures. However, contains some awkward phrasing ("Chloe didn't like rats before too") and unnatural expressions ("Many times, people say rats are dirty"). The meaning remains clear despite these issues.

Model GE: Balances simplicity with natural flow effectively. Uses appropriate A2 vocabulary while maintaining readability. Minor issue with "clever" which could be "smart" for consistency, but overall well-adapted.

Model MI: Uses appropriate vocabulary and sentence length for A2. Natural flow and accurate content. The phrase "she gets why" is slightly informal but understandable. Good balance of simplification and clarity.

["Model_LL", "Model_MI", "Model_GE", "Model_GP", "Model_CL", "Model_DE", "Model_QW"]