EhiMeNLP at TSAR 2025 Shared Task: Candidate Generation via Iterative Simplification and Reranking by Readability and Semantic Similarity

Rina Miyata[†] **Koki Horiguchi**[†] **Risa Kondo**[†] **Yuki Fujiwara**[‡] **Tomoyuki Kajiwara**[†*]
[†] Graduate School of Science and Engineering, Ehime University, Japan
[‡] Faculty of Engineering, Ehime University, Japan
[*] D3 Center, The University of Osaka, Japan
{miyata@ai., horiguchi@ai., kondo@ai., fujiwara@ai., kajiwara@} cs.ehime-u.ac.jp

Abstract

We introduce the EhiMeNLP submission, which won the TSAR 2025 Shared Task on Readability-Controlled Text Simplification. Our system employed a two-step strategy of candidate generation and reranking. For candidate generation, we simplified the given text into more readable versions by combining multiple large language models with prompts. Then, for reranking, we selected the best candidate by readability-based filtering and ranking based on semantic similarity to the original text.

1 Introduction

Text simplification (Alva-Manchego et al., 2020b) is a task of paraphrasing complex expressions into simpler ones while preserving the core meaning of a given text. This technology is utilized to support reading comprehension for diverse readers, including children (De Belder and Moens, 2010), language learners (Petersen and Ostendorf, 2007), and individuals with language impairments (Evans et al., 2014). Since reading ability varies significantly among individuals, recent studies on text simplification have focused on controlling readability (Scarton and Specia, 2018; Nishihara et al., 2019; Yanamoto et al., 2022; Agrawal and Carpuat, 2023; Barayan et al., 2025).

This paper describes our system submitted to TSAR 2025 shared task on readability-controlled text simplification (Alva-Manchego et al., 2025). Our system employs a two-step approach: first, generating candidates for simplified text with large language models (LLMs), and then reranking the candidates with embeddings. Our team was ranked first out of 20 teams in the official evaluation (AUTORANK). Our code is available on GitHub.[1]

2 TSAR 2025 Shared Task

In this competition (Alva-Manchego et al., 2025), participants are asked to simplify English paragraphs written at the Common European Framework of Reference for Languages (CEFR)[2] readability level of `B2` or more complex into simpler versions at levels `B1` and `A2`, respectively. The CEFR is the most widely used international standard for describing the language ability of English learners, consisting of six levels ranging from basic (`A1`) to proficient (`C2`).

The dataset provided consists of trial data, containing 40 paragraphs, and test data, containing 200 paragraphs. The output texts are automatically evaluated for both RMSE of readability and semantic similarity. Note that while both BERTScore (Zhang et al., 2020) and MeaningBERT (Beauchemin et al., 2023) are included in the official evaluation script for semantic similarity, only the latter is used in the official final ranking.

3 EhiMeNLP System

Figure 1 shows an overview of our system. Our system employs a two-step strategy of candidate generation and reranking; we describe these proposed methods in Sections 3.1 and 3.2. We then provide implementation details in Section 3.3. Finally, we report the results of preliminary experiments on the trial dataset in Section 3.4.

3.1 Step 1: Candidate Generation

We iteratively apply the proposed prompts shown in Figure 2 to multiple LLMs to generate simplification candidates. To diversify the candidates, we propose four types of prompts.

`P1`: fine-grained simplification To include simplified texts with various readability levels in our

[1] https://github.com/EhimeNLP/TSAR2025

[2] https://www.coe.int/en/web/common-european-framework-reference-languages

Proceedings of the Fourth Workshop on Text Simplification, Accessibility and Readability (TSAR 2025), pages 217–222

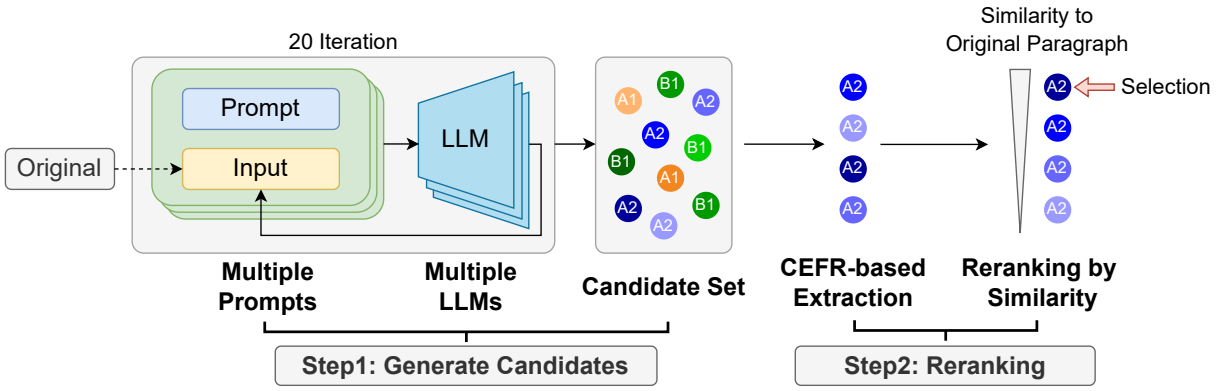

Figure 1: Overview of EhiMeNLP system.

candidate set, we iteratively generate paraphrases that are slightly simpler than the input text. Specifically, we define the readability level of the input text as i and instruct the LLMs to generate paraphrases at readability level $i - 1$. Note that this prompt does not provide any other explanation, such as a detailed description of the readability.

P2: controlling CEFR levels We employ the existing prompts (Barayan et al., 2025) that explicitly describe the reading comprehension abilities of learners at each CEFR level. Note that to adapt for this task, we modify the text unit from sentences to paragraphs. In addition, the target CEFR level changes incrementally from B2 to A1.

P3: controlling grade levels Instead of CEFR levels, we employ US grade levels, which are commonly used in previous studies on text simplification. In this prompt, we instruct the LLMs to simplify the input text to make it easier to understand for students in the target grade level. The grade level changes incrementally from 10 to 1.

P4: Clarifying edit operations A previous study (Kew et al., 2023) has reported the effectiveness of prompts that explicitly instruct editing operations (Alva-Manchego et al., 2020a) for LLM-based text simplification. We also utilize this prompt (Alva-Manchego et al., 2020a; Kew et al., 2023) to instruct LLMs to perform editing operations for text simplification, including paraphrasing, sentence compression, and sentence splitting.

3.2 Step 2: Reranking

In this step, we select the candidate with the highest semantic similarity to the input text while matching the target readability level.

Readability-based Filtering: First, we estimate the CEFR-based readability level for each candidate using the official evaluation script (Alva-Manchego et al., 2025). Then, we extract the set of candidates with the minimum difference from the target CEFR-based readability level.

Similarity-based Ranking: First, we estimate the semantic similarity between each candidate and the original text using the official evaluation script, based on both BERTScore[3] (Zhang et al., 2020) and MeaningBERT[4] (Beauchemin et al., 2023). Then, we select the candidate with the highest average score. In case multiple candidates achieve the highest average score, we select the candidate with the higher MeaningBERT score.

3.3 Implementation Details

As shown in Table 1, we employed six LLMs: `GPT-5`,[5] `GPT-4.1`,[6] `o3`,[7] `gpt-oss-20b`,[8] `Llama-3.3-70B-Instruct`,[9] and `Qwen3-32B`.[10] For `gpt-oss-20b`, we employed two configurations: one with `reasoning_effort` set to "low" and another with it set to "medium" to control the depth of thought. Regarding `Qwen3-32B`, we configured `enable_thinking`, which controls whether thinking occurs, to "False" from the perspective of inference speed.

[3] https://huggingface.co/spaces/evaluate-metric/bertscore
[4] https://huggingface.co/davebulaval/MeaningBERT
[5] https://platform.openai.com/docs/models/gpt-5
[6] https://platform.openai.com/docs/models/gpt-4.1
[7] https://platform.openai.com/docs/models/o3
[8] https://huggingface.co/openai/gpt-oss-20b
[9] https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct
[10] https://huggingface.co/Qwen/Qwen3-32B

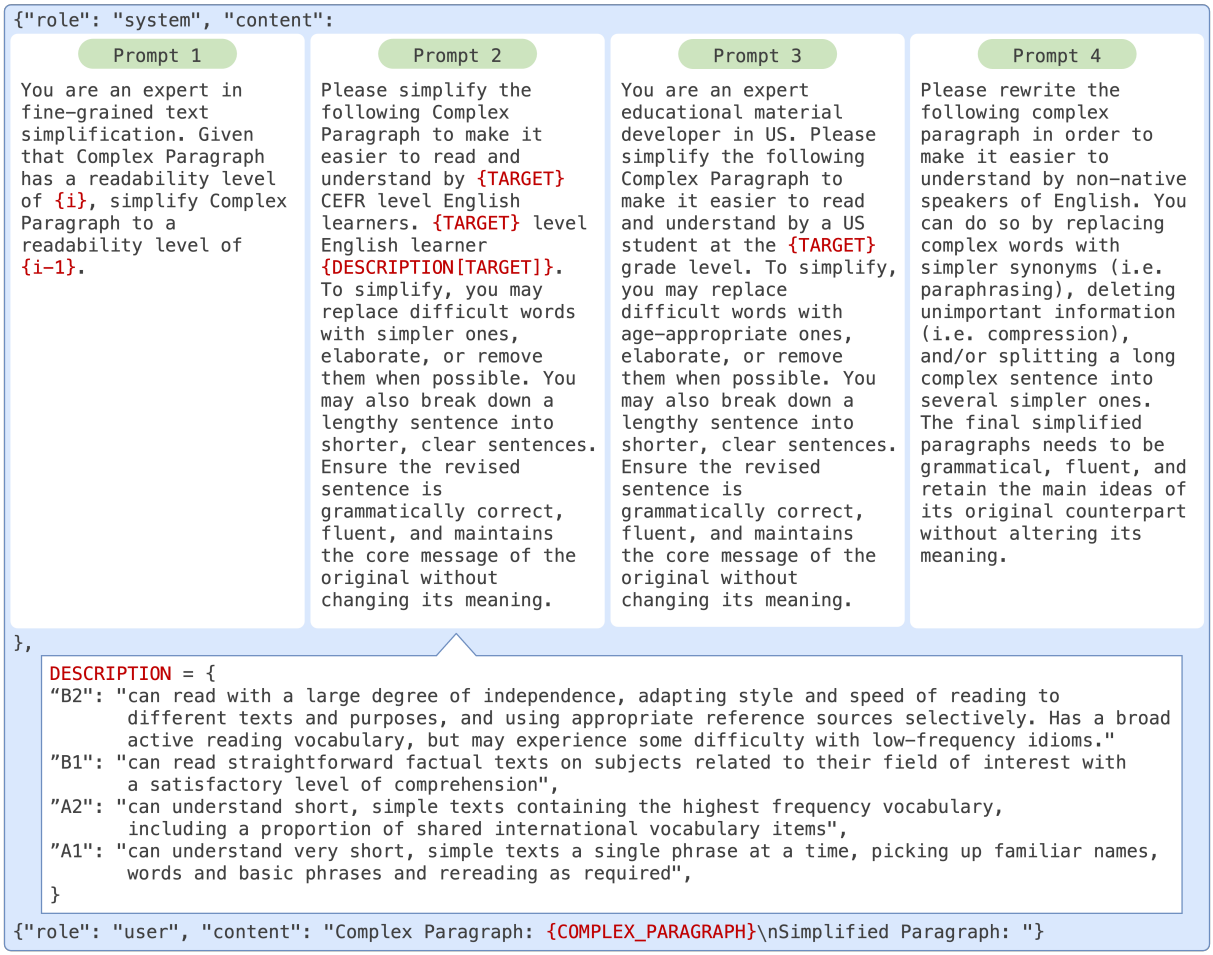

Figure 2: Prompts to generate simplification candidates.

All experiments were conducted using four RTX A6000 GPUs. For each LLM model, each prompt was run 20 times. Note that for prompt P2, there are 5 runs each for the 4 target CEFR levels, and for prompt P3, there are 2 runs each for the 10 target grade levels, totaling 20 runs in each case.

3.4 Preliminary Experiments

The left side of Table 2 shows the results of our preliminary experiments on the trial dataset. In this preliminary experiment, we applied four types of proposed prompts to seven types of LLMs to generate candidates using a total of 28 models, and counted how frequently each model was selected by our reranking. Experimental results reveal that the GPT-5 model is notably powerful and that the P1 prompt is remarkably useful.

Based on the results of the preliminary experiments, we have decided the three systems to be submitted as follows. Since the Llama-3.3-70B-Instruct model has only limited contributions, we decided not to employ it in our final system. In addition to the ensemble method for all LLMs and prompts, we decided to submit base models applying either the P1 prompt or the P3 prompt to the GPT-5 model.

4 Evaluation

Our EhiMeNLP system achieved first place in the official ranking (Alva-Manchego et al., 2025). As shown in Table 3, our system achieved a perfect score in the RMSE evaluation for readability. This demonstrates that our diverse set of candidates consistently generated text suitable for the target readability level. Regarding semantic similarity, our system achieved the fourth-highest score in both similarity to the source and reference texts, respectively. These results reveal that our system achieves readability control that balances both appropriate readability and high semantic similarity.

4.1 Ablation Analysis

Table 3 shows the performance of the base models, which apply the proposed prompts individually

Model	Reference	Inference	Token limit
GPT-5	gpt-5-2025-08-07	OpenAI API with greedy decoding	128,000
GPT-4.1	gpt-4.1-2025-04-14	OpenAI API with greedy decoding	32,768
o3	o3-2025-04-16	OpenAI API with greedy decoding	100,000
gpt-oss-20b	(OpenAI, 2025)	vLLM (Kwon et al., 2023)	40,000
Llama-3.3-70B-Instruct	(Llama Team, 2024)	vLLM (Kwon et al., 2023)	400
Qwen3-32B	(Qwen Team, 2025)	Transformers (Wolf et al., 2020)	32,768

Table 1: The LLM models used in this study.

	Trial					Test				
	P1	P2	P3	P4	Total	P1	P2	P3	P4	Total
GPT-5	6	2	4	2	14	30	19	12	6	67
Qwen3-32B	3	1	0	2	6	16	8	4	6	34
gpt-oss-20b (medium)	1	1	2	2	6	16	5	5	4	30
gpt-oss-20b (low)	3	0	1	1	5	22	3	5	2	32
GPT-4.1	3	0	1	0	4	12	8	2	1	23
o3	2	1	0	1	4	7	3	4	0	14
Llama-3.3-70B-Instruct	0	0	1	0	1	-	-	-	-	-
Total	18	5	9	8	40	103	46	32	19	200

Table 2: The frequency with which candidates generated by each model were finally selected.

Submission Name	Model	RMSE	MeaningBERT-orig	MeaningBERT-ref	Rank
EhiMeNLP / run1	Ensemble	0.000	0.902	0.845	1/48
EhiMeNLP / run2	GPT-5 with P1	0.200	0.838	0.816	7/48
-	GPT-5 with P2	0.265	0.850	0.836	-
EhiMeNLP / run3	GPT-5 with P3	0.234	0.847	0.840	6/48
-	GPT-5 with P4	0.394	0.844	0.836	-

Table 3: Results of the EhiMeNLP systems on the test dataset.

to the `GPT-5` model. The ensemble model outperformed the base models across all evaluation metrics. This highlights the importance of having a diverse set of candidates.

The `P1` prompt we submitted as our `run2` received relatively high scores for readability, but scored lower than other proposed prompts in terms of semantic similarity. The `P3` prompt we submitted as our `run3` outperformed `run2` in the official ranking due to its better balance between readability and semantic similarity.

4.2 Contributions of Each Model and Prompt

The right side of Table 2 shows how many times each combination of model and prompt was selected in the test dataset. Although prompt `P1` accounts for the majority, the other prompts also account for about half in total, indicating that combinations of multiple prompts are useful. In terms of LLM models, while `GPT-5` is the most frequently selected, `Qwen3-32B` and `gpt-oss-20b` also often appear, suggesting that combining multiple models contributes to improving the performance of the ensemble model.

5 Conclusion

We described the EhiMeNLP submission for the TSAR 2025 shared task. Our system employed a two-step strategy in which LLMs generated diverse candidates, followed by re-ranking based on readability and semantic similarity, achieving first place among 20 teams in the official ranking.

Lay Summary

This paper describes a text simplification system that paraphrases a given English text to a specific readability level. The TSAR 2025 workshop held a shared task on readability-controlled text simplification, with 20 teams competing to demonstrate the performance of their systems. Our EhiMeNLP system achieved the top performance among them.

Our system employed a two-step strategy: first, we leveraged large language models (LLMs) to generate diverse simplification candidates, and then selected the final output text through re-ranking based on readability and similarity. While LLMs are good at paraphrasing, they are not necessarily good at controlling readability levels. Therefore, we decided to generate a variety of paraphrases with different readability levels as candidates for simplification. To generate diverse candidates for simplification, we provided four types of prompts to six LLMs and performed repeated simplification. Our re-ranking step consists of two components: filtering based on readability and re-ranking based on similarity. This process enables our system to achieve high synonymity with the input text while respecting the target readability level.

According to official evaluations, our system perfectly satisfies the target readability while also achieving a high level of semantic similarity with both input and reference texts. Our detailed analysis revealed that while `GPT-5` is powerful, its ensemble with other LLMs is proving effective.

Acknowledgments

This work was supported by JST BOOST Program Japan Grant Number JPMJBY24036821, Cross-ministerial Strategic Innovation Promotion Program (SIP) on "Integrated Health Care System" Grant Number JPJ012425, and JSPS KAKENHI Grant Number JP25K03233.

References

Sweta Agrawal and Marine Carpuat. 2023. Controlling Pre-trained Language Models for Grade-Specific Text Simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12807–12819.

Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. 2020a. ASSET: A Dataset for Tuning and Evaluation of Sentence Simplification Models with Multiple Rewriting Transformations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4668–4679.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020b. Data-Driven Sentence Simplification: Survey and Benchmark. *Computational Linguistics*, 46(1):135–187.

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*.

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.

David Beauchemin, Horacio Saggion, and Richard Khoury. 2023. MeaningBERT: Assessing Meaning Preservation Between Sentences. *Frontiers in Artificial Intelligence*, 6.

Jan De Belder and Marie-Francine Moens. 2010. Text Simplification for Children. In *Proceedings of the SIGIR 2010 Workshop on Accessible Search Systems*, pages 19–26.

Richard Evans, Constantin Orăsan, and Iustin Dornescu. 2014. An evaluation of syntactic simplification rules for people with autism. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 131–140.

Tannon Kew, Alison Chi, Laura Vásquez-Rodríguez, Sweta Agrawal, Dennis Aumiller, Fernando Alva-Manchego, and Matthew Shardlow. 2023. BLESS: Benchmarking Large Language Models on Sentence Simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13291–13309.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, page 611–626.

Llama Team. 2024. The Llama 3 Herd of Models. *arXiv:2407.21783*.

Daiki Nishihara, Tomoyuki Kajiwara, and Yuki Arase. 2019. Controllable Text Simplification with Lexical Constraint Loss. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 260–266.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b Model Card. *arXiv:2508.10925*.

Sarah E. Petersen and Mari Ostendorf. 2007. Text Simplification for Language Learners: A Corpus Analysis. In *Proceedings of the 1st Workshop on Speech and Language Technology in Education*, pages 69–72.

Qwen Team. 2025. Qwen3 Technical Report. *arXiv:2505.09388*.

Carolina Scarton and Lucia Specia. 2018. Learning Simplifications for Specific Target Audiences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 712–718.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Daiki Yanamoto, Tomoki Ikawa, Tomoyuki Kajiwara, Takashi Ninomiya, Satoru Uchida, and Yuki Arase. 2022. Controllable Text Simplification with Deep Reinforcement Learning. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 398–404.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *Proceedings of the 8th International Conference on Learning Representations*.