OneNRC@TSAR2025 Shared Task: Small Models for Readability Controlled Text Simplification

Sowmya Vajjala
National Research Council, Canada
sowmya.vajjala@nrc-cnrc.gc.ca

Abstract

In this system description paper, we describe the team OneNRC's experiments on readability controlled text simplification, focused on using smaller, quantized language models ($< 20B$). We compare these with one large proprietary model and show that the smaller models offer comparable results in some experimental settings. The approach primarily comprises of an agentic workflow, and tool calling. The best results were achieved while using a CEFR proficiency classifier as a verification tool for the language model agent. In terms of comparison with other systems, our submission that used a quantized Gemma3:12B model that ran on a laptop achieved a rank of 9.88 among the submitted systems as per the AUTORANK framework used by the organizers. We hope these results will lead into further exploration on the usefulness of smaller models for text simplification.

1 Introduction

Automatic Text Simplification (ATS) is the task of translating a text written in a more advanced language into more accessible language. ATS research in NLP is over two decades old and most of the research focused on a single complex $->$ simple language text generation task, considering the unavailability of large scale graded simplification corpora. Siddharthan (2014) and Alva-Manchego et al. (2020) are two well-known surveys on the topic. The advent of Large Language Models (LLMs) made it possible to overcome the ATS dataset barrier to some extent with good zero-/few- shot performance on English text simplification (Kew et al., 2023) and some recent research explored zero-shot text simplification targeting specific reading levels (Farajidizaji et al., 2024; Barayan et al., 2025). Extending this strand of research, the TSAR 2025 Shared Task on Readability-Controlled Text Simplification (Alva-Manchego et al., 2025) aimed to compare systems that can generate simplified versions of text that conform to a target CEFR reading level, while preserving the meaning. No training data was provided and the test data had 100 texts, each with targeted simplification levels of A2 and B1 (i.e., 200 in total). Reference simplifications were provided after the task completion. AUTORANK (Kocmi et al., 2025) was used to rank the submitted systems.

Approach Summary: Team OneNRC's approach to this task focused on comparing how far can small language models go in zero-shot readability controlled ATS when they are supported with an agentic workflow and tool support. We used two tools: a) a CEFR prediction tool to help the language model agent verify its generated output and regenerate as needed, and b) an optional second tool to check the similarity between the generated text and the original text. For comparison, we also re-ran the same experiments with a larger proprietary model (Gemini-2.5-flash), and presented an analysis of the differences in the paper.

Motivation: Several small, open-weight language models were released with 1B-20B parameter range over the past year or two, and some of them can even be run for inference on consumer laptops, often in a quantized form. However, it is unclear how useful are they for many NLP tasks as they are not commonly compared with larger models in existing research. This gap motivates OneNRC's submission to this shared task.

In terms of the performance, one of the runs we submitted (using a 4-bit quantized Gemma3:12B model on a laptop) achieved an AUTORANK of 9.88, among the 40 submitted runs(Alva-Manchego et al., 2025). In this paper, we describe our approach in detail (Section 2), discuss other experiments which achieve better results than the submitted runs, share some qualitative observations (Section 3) along with broad conclusions on learnings through this study. (Section 4).

Proceedings of the Fourth Workshop on Text Simplification, Accessibility and Readability (TSAR 2025), pages 131–136
November 4-9, 2025 ©2025 Association for Computational Linguistics

2 Approach

Our approach primarily concentrated on exploring the power of small language models for this task, under the following zero-shot settings:

1. Plain prompting with local models
2. Prompting in a ReAct (Yao et al., 2022) agent setup using smolagents[1] with or without tool support. Two tools were used: a) a CEFR level prediction tool and b) a meaningBERT similarity score comparison tool - both provided by the organizers as a part of the shared task.
3. Prompting in a ReAct agent setup, but with structured prompt, input and output specification using dspy [2] (Khattab et al., 2024) instead of free-form prompting used in smolagents. The same two tools were explored in this setting too. Note that dspy supports further inbuilt automatic prompt optimization using few-shot examples, but we did not explore that in this paper.

These three settings focus on incrementally enhancing an LLM's built in capabilities without fine-tuning. The first setting relies on a hand-crafted prompt, and is potentially subject to more variation in the output depending on even minor changes in prompts. The ReAct approach aims to interleave reasoning and actions performed by the LLM with or without additional tools, and and is expected to provide more reliable outputs supported by a reasoning trace. Dspy focuses on structured specification of prompts, inputs, and outputs and emphasizes on a more programmatic interface to interacting with an LLM instead of handcrafted prompts, which further adds another layer of support to the base LLM. Thus, these three experimental settings can be viewed as a sequence of incremental adjustments to LLMs.

Compute and Costs: All experiments were run on a Macbook Pro M1 Pro laptop with a 32 GB RAM, and the costs for running the gemini-2.5-flash experiments amounted to 5 USD. All the implementation code including the used prompts are provided in the github for replication and reproduction.[3]

3 Results

We report results using the three official measures (RMSE, MeaningBERT-Orig, MeaningBERT-Ref) along with the weighted-F1 for CEFR level compliance (which our approach optimized for), and the following three models:

1. A 4-bit quantized Gemma3:12B (Team et al., 2025) model that does not natively offer tool support (which we nevertheless used with tool support).
2. GPT-oss-20B (Agarwal et al., 2025), a model post-trained with quantization of the model's mixture-of-expert weights to MXFP4 format [4], and can run natively without additional quantization on a laptop.
3. Gemini-2.5-Flash (Comanici et al., 2025), a large proprietary model as a comparison.

The first two models run on a consumer laptop using ollama[5]. The gemini-2.5-flash model was accessed through API calls with OpenRouter[6]. Our final results are summarized in Table 1.

Model	CEFR F1	RMSE	MB-Orig	MB-Ref
Basic Prompting with local models				
gemma	**0.6765**	**0.5745**	0.7514	0.7917
gptoss	0.5657	0.7517	**0.8216**	**0.8308**
ReAct Agent, No Tools				
gemma	**0.6585**	**0.6042**	0.7648	0.8055
gptoss	0.5612	0.7937	**0.8178**	**0.826**
gemini	0.6528	0.6124	0.7702	0.8152
ReAct Agent, CEFR Compliance Tool				
gemma	0.6559	0.6403	0.7304	0.7605
gptoss	0.7903	0.4796	**0.795**	**0.8152**
gemini	**0.9494**	**0.2449**	0.7689	0.8029
ReAct Agent, two tools				
gemma	0.6676	0.6671	0.7377	0.7616
gptoss	0.7969	0.4796	**0.7938**	**0.8177**
gemini	**0.917**	**0.3**	0.7749	0.8007
ReAct Agent in DSpy, No Tools				
gemma	0.622	0.6083	0.7466	0.7753
gptoss	0.5766	0.7616	**0.8209**	**0.8308**
gemini	**0.631**	**0.6**	0.7449	0.8027
ReAct Agent in DSpy, CEFR Compliance Tool				
gemma	0.7204	0.5292	0.7488	0.7844
gptoss	0.6071	0.7937	**0.8144**	**0.8243**
gemini	**0.9646**	**0.2**	0.7618	0.7997

Table 1: Comparison of results across different approaches

[1] https://smolagents.org/

[2] https://dspy.ai/

[3] https://github.com/nishkalavallabhi/tsarst2025-paper

[4] https://ollama.com/library/gpt-oss

[5] https://ollama.com/

[6] https://openrouter.ai/google/gemini-2.5-flash

Discarded models: We initially experimented with a some of the small (quantized) models that support tool use e.g., LLama3.2:1B and 3B, Qwen3:4B and 8B, primarily to start with as small models as possible, but adding tool support or agentic setup resulted in a drastic decline in performance for all models under 12B that we explored. For example, a react agent regenerated the same text for 20 iterations in one case, with the LLama3.2:1B model. So, we did not perform additional experiments with the models under 12B.

3.1 Discussion

Without any tool calling, in a plain prompt based setup, the quantized Gemma3 model is surprisingly better than a larger gpt-oss model in terms of CEFR compliance and there is no substantial difference compared to even a much larger Gemini model when no tools are used in a react agent setup. It is also important to note that there is not much difference between using a ReAct agent without tools versus just prompting the model across different evaluation measures. The local Gemma3 model without ReAct agent even shows slightly better CEFR compliance than the Gemini model in the same setup. While using a tool did not benefit Gemma3 much in a prompt + react agent setup, it did result in a 5% boost in CEFR compliance compared to basic prompt setup, when used with dspy, which could be attributed to the differences in the react prompting setup between smolagents and dspy, and dspy's focus on a more structured input approach to prompting. Considering that Gemma3 does not natively support tool use, it is interesting to note the performance improvement it could achieve with tool support and structured prompting. This could be further explored in future with few-shot prompt optimization. Gpt-oss and Gemini saw major improvements in overall CEFR compliance after tool support, but this came with a drop in both the meaningBERT similarity scores.

The usage of a second tool to verify meaning similarity did not seem to benefit any of the models, in any settings. Gpt-oss model consistently maintained better meaningBERT scores across all our experimental settings, and was consistently better than Gemma3 model even in CEFR compliance once tool calling was added to plain prompting. Clearly, adding tool calling benefited both in terms of CEFR compliance and meaning preservation.

Best results: Overall, the best result we achieved with a small model in terms of the official evaluation measures is with gpt-oss:20B (RMSE: 0.4796, MeaningBERT-Orig:0.7938, MeaningBERT-Ref:0.8177) in the ReAct+CEFR tool setup with smolagents, and the best result with Gemini-2.5-Flash is achieved in the ReACT+CEFR Tool+Dspy setting (RMSE: 0.2, MeaningBERT-Orig:0.7618, MeaningBERT-Ref:0.7997). For a comparison, the best submitted system amongst all the shared task systems (EHIMENLP) achieved the scores of 0 for RMSE, 0.902 and 0.845 for MeaningBERT-Orig and Ref respectively (Alva-Manchego et al., 2025). While we don't know the details of this system yet, it is interesting to see that high scores on meaningBERT were maintained despite low RMSE, which clearly was not the case in our experiments. In our experiments, we consistently observed that lower RMSEs are accompanied by lowering of the meaningBERT scores. Additionally, all our approaches had a higher MeaningBERT-Ref score than MeaningBERT-Orig score, unlike the EHIMENLP system.

Qualitative Comparison: To get a feel for how much difference do these approaches bring into the outputs themselves, Table 2 shows an example text and the outputs with the some of the approaches, along with the reference simplification. In this example, interestingly, except one (Gemini+React+2 Tools), all the approaches left out the expansion of the abbreviation. Gemma3 based approach appears to focus on generating short sentences. There is a small amount of hallucination in the gpt-oss model's output for this example, where "mother" in the original was transformed into "sister" in the simplified version. Only one model seems to miss the detail that that the asteroid is 800 meters wide, which is present in the original and reference texts. Going beyond these minor differences between models, the generated text generally seems to have undergone some form of simplification, and there are substantial similarities with the given reference text.

Table 3 shows an example text and a 4-bit quantized Gemma3 model's output for the two target CEFR levels-A2 and B1 using plain prompting without any tool support or agentic setup. The output shows that even this basic model is able to distinguish between the two levels to some extent in terms of language use. That indicates that smaller

Original	The first is DART – the Double Asteroid Redirection Test. Plans are scheduled to test DART on the moon of an asteroid called Didymos. 'Didymoon' is 150 metres wide, orbiting its 800-metre mother, and hopefully the impact of DART will knock it out of its orbit enough for Earth-based telescopes to pick up.
Plain Gemma3	There is a project called DART. DART will test something on the moon of an asteroid. The asteroid is called Didymos. Didymoon is part of Didymos. It is about 150 metres wide. Didymoon goes around Didymos, which is bigger - about 800 metres wide. We hope DART will push Didymoon a little bit. Then, telescopes on Earth will be able to see the change.
Gemma3+ ReAct + CEFR Tool	DART is a test. It will hit the moon of an asteroid called Didymos. The moon is called Didymoon. It is small, 150 meters. The asteroid is bigger, 800 meters. We want to see if DART can change Didymoon's path.
Gpt-oss + ReAct +CEFR Tool	DART is a mission to change an asteroid. Scientists will test it on the moon of the asteroid Didymos. The moon is called Didymoon. It is 150 metres wide. It goes around its larger sister, an asteroid 800 metres wide. The plan is that DART will hit Didymoon and move it so that telescopes on Earth can see it.
Gemini + ReAct + CEFR Tool	First, there's DART. It's a test to change an asteroid's path. We will test DART on Didymoon, a moon of an asteroid named Didymos. Didymoon is 150 meters wide. It moves around its parent asteroid. We hope DART will hit Didymoon and change its orbit a little. Then, telescopes on Earth can see this change.
Gemini + ReAct + 2 Tools	The first thing is DART (Double Asteroid Redirection Test). People want to test DART on a moon of an asteroid called Didymos. This small moon is called 'Didymoon'. It is 150 meters wide. It goes around its mother asteroid, which is 800 meters. We hope DART's hit will move Didymoon a little. Then telescopes on Earth can see it.
Reference	The first project is DART- the Double Asteroid Redirection Test. It will be tested on the moon of an asteroid called Didymos. This moon is called "Didymoon." It is 150 metres wide and goes around its bigger "mother" asteroid, which is 800 metres wide. Scientists hope that when DART hits Didymoon, it will change its path enough for telescopes on Earth to see.

Table 2: Example outputs across different systems for Textid:60-A2)

Original	Cape Town can be a dangerous place for baboons. Some have been injured or killed in car accidents and others have been electrocuted as they played with electric cables. Baboons are an endangered species and to save them and minimise conflicts between humans and baboons, the city council has employed a team of Baboon Monitors. The monitors' job is to chase baboons out of urban areas or catch them and release them in the countryside. This policy has had some success, though it is proving hard to stop baboons from coming back to the city once they see it as a source of easy food.
Plain Gemma3-A2	Baboons can be in danger in Cape Town. Sometimes they get hurt or die in car accidents. Sometimes they get hurt by electricity. Baboons are rare, so we need to protect them. The city has people called Baboon Monitors. Their job is to move baboons away from the city. They take them to the countryside. This helps, but it's hard to stop baboons from coming back because they find food easily in the city.
Plain Gemma3-B1	Baboons can face dangers in Cape Town. Sometimes they are hurt or die in car accidents, or get hurt by electricity when they play near wires. Baboons are rare, and we need to protect them. To help the baboons and avoid problems with people, the city has a team of Baboon Monitors. These monitors try to move baboons away from towns or catch them and take them to the countryside. This has helped a little, but it's difficult to stop baboons from returning to the city if they think they can easily find food there.
Reference	Cape Town can be dangerous for baboons. Some are hurt or die in car accidents. Some die from playing with electric wires. Baboons are animals that are in danger of disappearing. To help them, the city now has a team called Baboon Monitors. These people try to move baboons out of the city and take them back to the countryside. This plan works a little, but the baboons often come back because they find food in the city.

Table 3: Textid:25 in the A2 and B1 versions generated by the 4-bit quantized Gemma3:12B model, just using a simple prompt

language models may hold potential in generating simplifications and perhaps as writing support tools for creating simplified versions of text. Looking at multiple model generations may also be more informative for human writers.

4 Conclusions

Overall, our experiments lead us to three main conclusions:

1. Small language models are capable of doing some readability controlled text simplification even in a zero-shot setting.
2. Supplying appropriate tool calls is useful, especially for the larger models.
3. There appears to be a tradeoff between achieving the right CEFR level and preserving the meaning.

Evaluation: Evaluation in this shared task primarily relied on automated models, and what appears like an ad-hoc weighting scheme. While a smaller scale human evaluation may be needed to understand the generated texts better, explanatory measures using LLM-as-a-Judge approaches or checklist based evaluation approaches (Cook et al.; Mohammadkhani and Beigy, 2025) may provide more informative evaluation for text simplification. Small language models may hold a potential as judges too, providing low-cost, low-carbon footprint options. Finally, manual analysis revealed the possibility of hallucination even in these constrained generation scenarios. As the use of LLMs for text simplification increases, the evaluation measures need to also account for hallucination as a potential dimension.

Future Work: While a laptop environment is not conducive to conduct experiments on a larger scale, we believe future work should focus on exploring the power of smaller models in few-shot settings or fine-tuning using synthetic data generation. Dspy's automatic prompt optimization capabilities in few shot settings should be better explored for this task, to understand the true potential of smaller models for text simplification. The current shared task focused only on English, but these experiments lead us to hypothesize that there may be some possibilities to replicate these results for at least some of the high resource languages, which needs to be explored. Finally, using language models as support tools to aid the authors of simplified texts also seems to be a useful direction to pursue.

Limitations

We did not venture into few-shot prompting or any other in-context learning strategies in this paper. Further, since everything was ran on a laptop, we only used 4-bit quantized versions of several small models, and their unquantized versions are potentially more powerful. This can be explored further in future, and the current results should be considered under these limitations.

Lay Summary

Automatic Text Simplification (ATS) is the task of translating a text written in a more advanced language into more accessible language. A common approach to solve this problem is to collect large number of examples of sentences and their manually simplified variants (e.g., from some form of sentence level alignment of Wikipedia and Simple Wikipedia articles) and use them as inputs to computational algorithms that are capable of learning patterns from such large amounts of data. The output of this process creates a simplification "model", which can be used to simplify any given new text. One way to add more nuance to this process is to instruct the whole process to do graded simplification, based on some scale such as the CEFR scale, which is used to describe language proficiency at various levels. However, that would also mean collecting large amounts of examples for each level. Hence, this kind of research was largely restricted to languages where such data is already available or can be collected easily.

The arrival of general purpose Artificial Intelligence based models like ChatGPT gave some flexibility to this approach. We can now describe our problem, potentially give a few examples, and prompt a large language model to simplify a given text, targeting a specific CEFR level. This shared task (where different groups of researchers work on solving a given problem under the same data/evaluation conditions) focuses on CEFR targeted text simplification for English. The approach described in this paper explores whether we really need very large models for this task, asking the question: what can small language models that can run on laptops achieve, if they are given some support in the form of tools that can verify their output and tools that can automatically adapt the human written prompts to suit the language model's requirements. It turns out, they can put up a strong competition to larger ones, sometimes.

References

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, and 1 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020. 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 (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.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Jonathan Cook, Tim Rocktäschel, Jakob Nicolaus Foerster, Dennis Aumiller, and Alex Wang. Ticking all the boxes: Generated checklists improve llm evaluation and generation. In *Language Gamification-NeurIPS 2024 Workshop*.

Asma Farajidizaji, Vatsal Raina, and Mark Gales. 2024. Is it possible to modify text to a target readability level? an initial investigation using zero-shot large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9325–9339.

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.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Saiful Haq, Ashutosh Sharma, Thomas T Joshi, Hanna Moazam, Heather Miller, and 1 others. 2024. Dspy: Compiling declarative language model calls into state-of-the-art pipelines. In *The Twelfth International Conference on Learning Representations*.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Natalia Fedorova, Mark Fishel, Markus Freitag, and 1 others. 2025. Preliminary ranking of wmt25 general machine translation systems. *arXiv preprint arXiv:2508.14909*.

Mohammad Ghiasvand Mohammadkhani and Hamid Beigy. 2025. Checklist engineering empowers multilingual llm judges. *arXiv preprint arXiv:2507.06774*.

Advaith Siddharthan. 2014. A survey of research on text simplification. *ITL-International Journal of Applied Linguistics*, 165(2):259–298.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *NeurIPS 2022 Foundation Models for Decision Making Workshop*.