

HULAT-UC3M at TSAR 2025 Shared Task: A Prompt-Based Approach using Lightweight Language Models for Readability-Controlled Text Simplification

Jesus M. Sanchez-Gomez¹, Lourdes Moreno¹
Paloma Martínez¹, Marco Antonio Sanchez-Escudero¹
{jesusmsa, lmoreno, pmf, marcoasa}@inf.uc3m.es

¹Universidad Carlos III de Madrid, Av. Universidad, 30, Leganés, 28911, Spain

Correspondence: marcoasa@inf.uc3m.es

Abstract

This paper describes the participation of the HULAT-UC3M team in the TSAR 2025 Shared Task on Readability-Controlled Text Simplification. Our approach uses open and lightweight Large Language Models (LLMs) with different sizes, together with two strategies for prompt engineering. The proposed system has been tested on the trial data provided, and evaluated using the official metrics: CEFR Compliance, Meaning Preservation, and Similarity to References. LLaMA 3 8B model with reinforced prompts was selected as our final proposal for submission, and ranking fourteenth according to the overall metric. Finally, we discuss the main challenges that we identified in developing our approach for this task.

1 Introduction

Automatic text simplification aims to generate texts that are easier to read and understand while preserving their original meaning (Saggion and Hirst, 2017). Large Language Models (LLMs) are increasingly being explored for this purpose adapting generated content to match users' specific language proficiency levels.

The Common European Framework of Reference for Languages (CEFR) grades language mastery on a six-level scale, ranging from Basic User (A1, A2), through Independent User (B1, B2), to Proficient User (C1, C2). Despite the promise of LLMs for generating simplified texts, accurately classifying these outputs into specific CEFR levels remains a significant challenge. Current works indicate that LLMs often produce unsatisfactory results in CEFR-related educational tasks, frequently overpredicting intermediate levels like B1 or B2 and struggling to consistently encode or apply intrinsic CEFR knowledge without explicit prompting (Benedetto et al., 2025). This highlights the need for improved automatic evaluation metrics specifically tailored for Readability-Controlled Text Sim-

plification (RCTS), as standard metrics often misinterpret simplification operations and inaccurately assess readability and meaning preservation.

HULAT-UC3M participation in TSAR 2025 Shared Task (Alva-Manchego et al., 2025) integrates generative LLMs through prompt engineering by testing open lightweight LLMs of different sizes using prompts with different levels of detail.

2 Related Work

The challenge of Readability-Controlled Text Simplification involves rewriting text to match specific audience proficiency levels, most commonly guided by the CEFR. Recent research has evidenced a concentrated effort to leverage the power of LLMs to achieve this fine-grained control (Agrawal and Carpuat, 2023). For instance, studies have investigated the performance of pre-trained instruction-tuned LLMs in aligning their generated output with standardized readability criteria, finding that models vary widely in their inherent ability accurately to encode and apply CEFR requirements when performing generative tasks (Imperial and Madabushi, 2023; Benedetto et al., 2025). In addition, methods have been developed to integrate readability control directly into the generation process itself, including techniques that use reinforcement learning and decoding approaches with anticipation mechanisms to minimize the gap between the requested and observed readability levels (Ribeiro et al., 2023).

A key methodological advancement in RCTS has been the transition towards controlling complexity via prompt-based strategies. Many recent approaches focus on the zero-shot or few-shot prompting, where pre-trained models are guided entirely by explicit or a small number of in-context examples to a target complexity level (Farajidizaji et al., 2024; Chi et al., 2023). These strategies allow modifying any source text to an absolute target read-

ability level, often making the generated output less dependent on the complexity of the original source. The effectiveness of these techniques highlights the role of prompt richness, the degree of descriptive guidance provided, in controlling readability. This aspect is directly connected to our approach, which compares reinforced and slightly reinforced prompts as two levels of guidance for readability-controlled simplification.

We conclude that, given the strong performance observed in recent literature, our approach will focus exclusively on zero-shot strategies for pre-trained LLMs, leveraging only advanced prompt engineering to control text simplification. This methodology avoids the need for exhaustive fine-tuning on domain-specific corpora, allowing a highly flexible and efficient solution.

3 Proposed Architecture

Two families of LLMs have been considered: Ettin suite (Weller et al., 2025b) and Meta LLaMA 3 (Grattafiori et al., 2024). As for prompt engineering, the reinforced prompts have consisted of the description of each CEFR level in order to achieve the simplification of the original text to the required target CEFR level. Figure 1 shows the architecture of our approach. After reviewing the latest advances on recent, open, lightweight, and task-suitable LLMs, LLaMA 3 and Ettin Decoder are selected as generation modules. The prompt engineering techniques are developed for being applied with the adapted and CEFR-aligned datasets. This key step involves the use of reinforced prompt and slightly reinforced prompt to guide the models’ text generation. Afterwards, the approach is assessed in the automatic evaluation metrics module, evaluated using the official set of evaluation metrics and our complementary ones. A manual review is also carried out in the form of a qualitative inspection.

Finally, after quantitative and qualitative analysis, a decision is made to select the final configuration for presentation. The final configuration was selected prioritizing CEFR Compliance and Align-Score, supported by manual qualitative inspection.

3.1 Large Language Models used

On the one hand, LLaMA 3 herd of models (Grattafiori et al., 2024) includes two versions of pre-trained and instruction-tuned generative text models with sizes of 8B and 70B AI@Meta (2024). For the purpose of using lightweight models, the

version with 8B parameters has been chosen for our proposal.

On the other hand, Ettin Suite of models (Weller et al., 2025b) is a novel collection of state-of-the-art paired encoder-only and decoder-only language models, ranging from 17M to 1B parameters publicly available in Weller et al. (2025a). Considering the results obtained by these models in the different tasks, Ettin Decoder version with a size of 400M parameters has been chosen for our approach, which will allow us to make comparisons between very lightweight open-source models and small public models.

As for the hyperparameters used, we adopted slightly different configurations for each model. In the case of LLaMA 3, we used: temperature = 0.3, top-p = 0.9, and no. of tokens = 256. For Ettin Decoder, we followed the official guidelines, and used temperature = 0.7, top-p = 0.9, and limiting the no. of tokens between 100 and 256.

3.2 Prompt Design

The prompt engineering technique consisted of developing reinforced prompts based on those used in Benedetto et al. (2025) for classification into CEFR levels. First, a reinforced prompt that includes detailed CEFR descriptions, in addition to examples for each of the six levels (from A1 to C2) in a few-shot approach. And second, a slightly reinforced prompt, which has consisted of shorter CEFR descriptions, with a more concise style, specifying only the simplest levels (A1, A2, and B1). These prompts are included in Appendix A, and are publicly available in the HULAT-UC3M GitHub¹ repository (Human Language and Accessibility Technologies (HULAT) group, 2025).

4 Experimental Setup and Evaluation

4.1 Dataset

The trial data provided contained 20 original texts, and for each of them, both their simplification to CEFR level A2 and level B1 have been included. The texts cover a diverse range of general knowledge and real-world topics, making this dataset suitable for a broad evaluation of a model’s ability to simplify different types of content. Regarding the test data provided, it has consisted of 100 original texts about similar topics, and simplifications to both CEFR level A2 and level B1 have been asked.

¹https://github.com/hulat-group/tsar_2025_workshop/blob/main/README.md

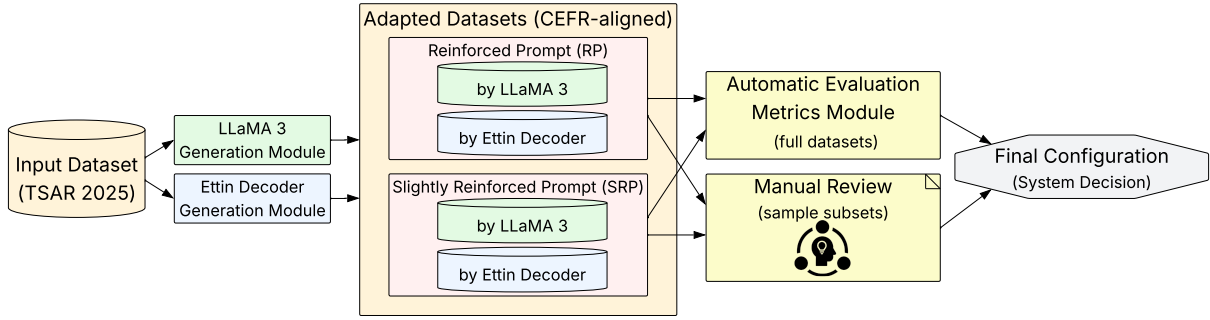


Figure 1: Flowchart of the architecture of the proposed approach.

4.2 Evaluation Metrics

We used the official evaluation metrics of the shared task: CEFR Compliance, Meaning Preservation, and Similarity to References. CEFR Compliance evaluates how well the generated text matches the requested CEFR level, using models fine-tuned on subsets of the Universal CEFR dataset (Imperial et al., 2025) and assessed through Weighted F1, Adjacent Accuracy, and RMSE (Barayan et al., 2025). Meaning Preservation is computed with MeaningBERT (Beauchemin et al., 2023), a BERT-based metric trained to correlate strongly with human judgments of semantic similarity. Similarity to References is evaluated with BERTScore (Zhang et al., 2020), which leverages contextual embeddings to estimate token-level similarity between candidate and reference texts, although it may overestimate lexical overlap without ensuring true semantic preservation.

In addition, we incorporated AlignScore (Zha et al., 2023) to assess factual consistency between the original and simplified texts. This complementary metric addresses the tendency of reference-based measures to overvalue surface similarity even when outputs include factual errors or hallucinations. AlignScore proved particularly valuable in revealing limitations of the Ettin Decoder model, whose results appeared adequate under official metrics but showed inconsistencies when factual alignment was evaluated.

5 Results and Discussion

5.1 Internal Evaluation on Trial Data

We conducted an internal evaluation using the trial data provided to compare our configurations before the final submission. Table 1 presents the results obtained in these experiments. Overall, the configurations using slightly reinforced prompts (SRP) yielded a better trade-off between readabil-

ity control and meaning preservation, whereas fully reinforced prompts (RP) led to higher variance and reduced factual consistency.

LLaMA 3 8B models. Among all systems tested, the LLaMA 3 8B model with a slightly reinforced prompt was the top performer, achieving the highest scores in CEFR Compliance and Similarity to References. This suggests that overly detailed prompts may hinder instruction-tuned models such as LLaMA 3, reducing their ability to generalize.

Ettin Decoder 400M models. For the Ettin Decoder 400M model, the RP version achieved competitive results in CEFR Compliance and strong BERTScore values for Meaning Preservation, but underperformed in the remaining metrics. The SRP version showed very high scores in Meaning Preservation but weak performance in CEFR Compliance. In terms of the complementary AlignScore metric, Ettin Decoder 400M obtained the lowest factual consistency (0.5576), despite its strong semantic similarity metrics, revealing inconsistent and unreliable outputs.

A qualitative inspection of the texts produced by Ettin Decoder 400M revealed frequent hallucinations and repetitions, making its outputs less usable despite high similarity scores. In contrast, the LLaMA 3 8B model achieved the highest AlignScore (0.6038), confirming it as the most balanced configuration between Meaning Preservation and factual fidelity. Therefore, for the final submission, we selected the LLaMA 3 8B model guided by a slightly reinforced prompt.

5.2 Error Analysis

This section provides an in-depth qualitative analysis of errors observed in the texts generated with the trial data by LLaMA 3 and Ettin Decoder models. Errors were classified into the following categories: hallucination; little or no simplification;

Metric	LLaMA 3 8B (RP)	LLaMA 3 8B (SRP)	Ettin Decoder 400M (RP)	Ettin Decoder 400M (SRP)
CEFR Compliance - Weighted F1	0.3000	0.5200	0.4800	0.2800
CEFR Compliance - RMSE	1.1100	0.7746	0.8900	1.1619
Meaning Preservation - MeaningBERT	0.6532	0.7170	0.6901	0.8176
Meaning Preservation - BERTScore	0.8837	0.8999	0.9025	0.8807
Similarity to References - MeaningBERT	0.6384	0.7075	0.6243	0.6199
Similarity to References - BERTScore	0.8764	0.8921	0.8789	0.8470
AlignScore	0.5600	0.6038	0.4300	0.5576

Table 1: Results of experiments with trial data (RP means reinforced prompt, and SRP slightly reinforced prompt).

repetitions; incoherent or low-fluency output; over-simplification and loss of nuance; rigid, artificial, or robotic output; and overloaded output with unnecessary details. Tables 3 and 4 in Appendix B include at least one example for each category.

Ettin Decoder models. The Ettin Decoder model (both RP and SRP) exhibited a higher frequency of severe output quality issues, with hallucination being the most critical error type (Table 3). In example #1, the RP version completely diverges from the original text about NASA tracking asteroids, generating content about a “new space telescope.” A similar hallucination occurs in the SRP version for the business etiquette example #5, where the model produces a list of fabricated, generic points with no semantic link to the source.

Another recurrent error type was low-fluency or repetitive output. In the RP examples #3 and #4, the model tends to loop, as in the supermarket chains example, where it repeats the same sentence multiple times, reducing readability. This behavior likely stems from issues with stopping criteria or reinforcement signals.

In the SRP example #7, based on *The Life of Pi*, the model adds a redundant and factually inaccurate detail about the boy being a vegetarian who must feed the tiger information absent from the original text again resulting in incoherent output. The “little or no simplification” error was also common across both versions, as illustrated by examples #2 and #6 on earthquake-proof bridges and asteroids. This suggests that, at higher input complexity levels, the model tends to reproduce the source text almost verbatim.

Finally, a recurring issue in the SRP version is that its outputs, while fluent and grammatically correct, are often overly simplified and lack nuance. As illustrated in example #8 concerning potential causes of bridge collapses, the model compresses the content into a generic summary, losing the explanatory richness of the original and particularly the nuances regarding design limitations and inves-

tigation processes.

LLaMA 3 models. As for the LLaMA 3 model (Table 4), although it generally produces more consistent outputs than Ettin Decoder, it is often affected by unnecessary meta-commentary, leading to rigid or overly formulaic text. The examples #9 and #11, related to *The Hunger Games*, illustrate this problem: the simplified text is followed by a generic note unrelated to the simplification task, likely a remnant of instruction-tuning data.

The error of over-simplification and loss of nuance is evident in example #10 for RP version about the five-day work week. Although the essential information is preserved, the simplification diminishes the narrators emotional tone and reduces the subjective depth of the original text. Regarding the SRP version, it also exhibited issues with unnecessary details. In example #12 about *Wasfia Nazreen*, the model simplifies the text but includes the explicit reason for the hula-hooping in a way disconnected from the narrative flow. Furthermore, it often breaks simple ideas into short, choppy sentences (e.g., “Wasfia just watched”), making the output redundant and stylistically uneven.

A common pattern observed across models was overshooting. The outputs displayed a consistent tendency to exceed the target CEFR level that is, when an A2 level was requested, the simplification was frequently evaluated as B1. This discrepancy aligns with the findings of [Benedetto et al. \(2025\)](#), who identified a bias of LLMs toward intermediate levels (B1/B2). Despite sharing this bias, the SRP versions outperformed the RP ones, reducing RMSE and improving both Meaning Preservation and AlignScore.

In summary, the errors in LLaMA 3 are primarily stylistic and structural, largely stemming from residual artifacts of instruction alignment, whereas Ettin Decoder exhibits more fundamental problems involving consistency, repetition, and factual integrity.

5.3 Evaluation on Test Blind

To generate the simplified texts from these test data, we have followed the same procedure as in the experiments with trial data. Following the release of the task ranking (Alva-Manchego et al., 2025), we have been able to analyze the performance achieved by our proposals in the test data. In addition to the set evaluation metrics used in the task, the AUTORANK score (Kocmi et al., 2025) has been used as an overall metric for the final classification.

Table 2 shows the official results of our approach in TSAR 2025 (the position is indicated over 48 submitted runs). Based on these results, the LLaMA 3 8B model with slightly reinforced prompt consistently outperforms the version with reinforced prompt. It achieves a significantly lower RMSE and a higher AvgScore, indicating more accurate predictions on average. The MeaningBERT-orig and MeaningBERT-ref scores are also higher, confirming its superior ability to preserve the original meaning while producing text that is more similar to human-written references. As for the final ranking position, SRP version ranked 24th, while RP version ranked 31st. Overall, the use of a slightly reinforced prompt appears to be a key factor in improving the model’s performance in this task.

5.4 Discussion

Our experiments highlight several key challenges in readability-controlled text simplification. The primary difficulty lies in generating outputs that accurately match specific CEFR levels: although LLMs encode information about the framework, they still struggle to apply it consistently in generative tasks. Another major challenge concerns managing the well-known trade-off between improving readability and preserving the original meaning.

In our internal evaluation with trial data, we compared recent and lightweight LLMs LLaMA 3 and the Ettin suite focusing on decoder models given the generative nature of the task. Despite the simplicity of our setup, the goal was to test whether a minimal prompt design could perform comparably to a reinforced one. Contrary to expectations, lighter prompts produced better results than more detailed ones, contrasting with previous findings such as Benedetto et al. (2025), who reported that instruction-tuned models required explicit CEFR descriptors for optimal performance.

With LLaMA 3 8B, the slightly reinforced

prompt consistently outperformed the reinforced variant. We attribute this to the models strong instruction-tuning and the high quality of the TSAR dataset, which reduce the need for extensive prompt engineering. This finding refines our understanding of level control in LLMs: in instruction-tuned architectures, concise and well-structured prompts can be more effective than exhaustive descriptions.

Regarding the Ettin Decoder, the qualitative analysis revealed substantial issues outputs were often incoherent or nonsensical despite claims of competitive performance in other benchmarks. Consequently, no test submission was made for this model. This outcome underscores the importance of manual inspection, which remains crucial to identify alignment and factuality problems that automated metrics may overlook.

In the blind test evaluation, our best system ranked 14th out of 20 teams by AUTORANK, achieving moderate scores in CEFR Compliance and Meaning Preservation (MeaningBERT-orig and -ref). These results reflect the inherent tension between the two dimensions of the task: controlling readability while maintaining semantic fidelity.

Looking ahead, improving CEFR-level control will likely require larger, dedicated CEFR-annotated corpora for fine-tuning. From an evaluation perspective, future tasks could benefit from robust factuality-oriented metrics such as AlignScore, trained on millions of cross-domain examples. Finally, a hybrid architecture combining an encoder specialized in CEFR assessment with a decoder optimized for controlled text generation could better satisfy the objectives of readability-controlled simplification.

Acknowledgments

This work has been supported by grant PID2023-148577OB-C21 (Human-Centered AI: User-Driven Adapted Language Models-HUMAN_AI) by MICIU/AEI/ 10.13039/501100011033 and by FEDER/UE.

Lay Summary

People often read things that are too complicated to understand quickly. This frequently happens with news articles, legal documents, or scientific papers. To solve this problem, researchers use computer programs called Large Language Models (LLMs) to automatically simplify text. LLMs are programs that can create human-like language. The main goal

Model	Position	RMSE	MeaningBERT-orig	MeaningBERT-ref	AvgScore	AUTORANK
LLaMA 3 8B (RP)	31	0.682	0.790	0.791	-0.122	10.560
LLaMA 3 8B (SRP)	24	0.608	0.793	0.806	-0.028	9.370

Table 2: Results obtained in TSAR 2025 (RP means reinforced prompt, and SRP slightly reinforced prompt).

is to make a text much easier to read without changing its main message. One way to measure whether a text is simpler is by using readability levels, such as the CEFR (Common European Framework of Reference). CEFR uses levels like A1 (beginner) to C2 (expert) to rate language skills. Our goal was to train these programs to simplify the text precisely to these specific reading levels.

We wanted to know how well language models could simplify text to specific reading levels. In particular, could we use simple instructions to get these models to generate simplified text that (1) keeps the original meaning, (2) achieve the desired reading level, and (3) avoid errors? Answering this question is important to make text more accessible to everyone.

We tested several small and open-source LLMs, including the LLaMA 3 and Etti Decoder, on text simplification tasks. We found that giving the models detailed instructions, known as the prompt-based approach, helped them keep the text meaning of the text very well. However, this approach failed to produce text that was easy to read. The models often made mistakes, such as creating rigid and robotic text, or including not required details. LLaMA 3 was the best, but current tools still fail to decide if a simplified text is natural and clear.

At last, this research can help many people who have difficulty reading complex texts. This includes students who are learning a new language, people with cognitive impairments, or citizens who need to understand legal or medical documents.

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. Association for Computational Linguistics.
- AI@Meta. 2024. [Llama 3 Model Card](#).
- 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*, volume 1, pages 6762–6781. Association for Computational Linguistic.
- David Beauchemin, Horacio Saggion, and Richard Khoury. 2023. [MeaningBERT: assessing meaning preservation between sentences](#). *Frontiers in Artificial Intelligence*, 6:1223924.
- Luca Benedetto, Gabrielle Gaudeau, Andrew Caines, and Paula Buttery. 2025. [Assessing how accurately large language models encode and apply the common European framework of reference for languages](#). *Computers and Education: Artificial Intelligence*, 8:100353.
- Alison Chi, Li-Kuang Chen, Yi-Chen Chang, Shu-Hui Lee, and Jason S Chang. 2023. [Learning to Paraphrase Sentences to Different Complexity Levels](#). *Transactions of the Association for Computational Linguistics*, 11:1332–1354.
- 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. ELRA Language Resource Association.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. [The Llama 3 Herd of Models](#). *arXiv preprint arXiv:2407.21783*.
- Human Language and Accessibility Technologies (HULAT) group. 2025. [HULAT-UC3M @ TSAR 2025 Shared Task](#).
- Joseph Marvin Imperial, Abdullah Barayan, Regina Stodden, Rodrigo Wilkens, Ricardo Munoz Sanchez, Lingyun Gao, Melissa Torgbi, Dawn Knight, Gail Forey, Reka R. Jablonkai, and 1 others. 2025. [UniversalCEFR: Enabling Open Multilingual Research on Language Proficiency Assessment](#). *arXiv preprint arXiv:2506.01419*.

Joseph Marvin Imperial and Harish Tayyar Madabushi. 2023. [Flesch or Fumble? Evaluating Readability Standard Alignment of Instruction-Tuned Language Models](#). In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 205–223. Association for Computational Linguistics.

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*.

Leonardo FR Ribeiro, Mohit Bansal, and Markus Dreyer. 2023. [Generating Summaries with Controllable Readability Levels](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11669–11687. Association for Computational Linguistics.

Horacio Saggion and Graeme Hirst. 2017. [Automatic Text Simplification](#), volume 32. Springer.

Orion Weller, Kathryn Ricci, Marc Marone, Antoine Chaffin, Dawn Lawrie, and Benjamin Van Durme. 2025a. [Ettin Suite: SoTA Paired Encoders and Decoders](#).

Orion Weller, Kathryn Ricci, Marc Marone, Antoine Chaffin, Dawn Lawrie, and Benjamin Van Durme. 2025b. [Seq vs Seq: An Open Suite of Paired Encoders and Decoders](#). *arXiv preprint arXiv:2507.11412*.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. [AlignScore: Evaluating Factual Consistency with A Unified Alignment Function](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 11328–11348. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [BERTScore: Evaluating Text Generation with BERT](#). In *International Conference on Learning Representations*.

A Prompts used

The full prompts used for the lightweight LLMs (see Subsection 3.1), reinforced and a slightly reinforced, are detailed below.

A.1 Reinforced Prompt

=====
PROMPT FOR CEFR LEVEL A1 (REINFORCED)

Below are CEFR descriptions:

- A1: Beginner Simple sentences. No passive. Top 1000 words. Familiar names, cognates.

- A2: Elementary Short, simple texts with predictable info. 1000–2000 words.
- B1: Intermediate Everyday or job-related language. Includes descriptions of events, feelings and wishes.
- B2: Upper Intermediate Contemporary articles, styles, viewpoints. 5000–10,000 words.
- C1: Proficient Long, complex factual/literary texts. 10,000–20,000 words.
- C2: Advanced Proficient All forms of written language including abstract and highly complex.

Example A1:

Original: Some asteroids are very small.

Simplified: Some space rocks are very small.

Target CEFR level: A1

Now simplify this one:

Original: {INPUT}

=====
PROMPT FOR CEFR LEVEL A2 (REINFORCED)

Below are CEFR descriptions:

- A1: Beginner Simple sentences. No passive. Top 1000 words. Familiar names, cognates.
- A2: Elementary Short, simple texts with predictable info. 1000–2000 words.
- B1: Intermediate Everyday or job-related language. Includes descriptions of events, feelings and wishes.
- B2: Upper Intermediate Contemporary articles, styles, viewpoints. 5000–10,000 words.
- C1: Proficient Long, complex factual/literary texts. 10,000–20,000 words.
- C2: Advanced Proficient All forms of written language including abstract and highly complex.

Example A2:

Original: Earthquakes damage buildings and bridges.

Simplified: Earthquakes can break buildings and bridges.

Target CEFR level: A2

Now simplify this one:

Original: {INPUT}

=====
PROMPT FOR CEFR LEVEL B1 (REINFORCED)

Below are CEFR descriptions:

- A1: Beginner Simple sentences. No passive. Top 1000 words. Familiar names, cognates.
- A2: Elementary Short, simple texts with predictable info. 1000–2000 words.
- B1: Intermediate Everyday or job-related language. Includes descriptions of events, feelings and wishes.
- B2: Upper Intermediate Contemporary articles, styles, viewpoints. 5000–10,000 words.
- C1: Proficient Long, complex factual/literary texts. 10,000–20,000 words.
- C2: Advanced Proficient All forms of written language including abstract and highly complex.

Example B1:

Original: Many wild animals are starting to enter cities.

Simplified: Some wild animals are visiting cities now.

Target CEFR level: B1

Now simplify this one:

Original: {INPUT}

A.2 Slightly Reinforced Prompt

PROMPT FOR CEFR LEVEL A1

Rewrite the following text at CEFR level A1.

Use very short and simple sentences. Only basic vocabulary. Avoid complex grammar.

Original: {INPUT}

Simplified:

PROMPT FOR CEFR LEVEL A2

Rewrite the following text at CEFR level A2.

Use short, simple sentences and clear, everyday words.

Original: {INPUT}

Simplified:

PROMPT FOR CEFR LEVEL B1

Rewrite the following text at CEFR level B1.

Use clear, connected sentences with some detail. Everyday vocabulary is fine, but avoid very advanced or abstract terms.

Original: {INPUT}

Simplified:

B Examples for Error Analysis

Tables 3 and 4 show the analysis of several examples of errors obtained in the generation of simplified text from trial data by Ettin Decoder and by LLaMA 3 models, respectively. These errors have been used and explained in Subsection 5.2.

Model	Id.	Original text	Simplified text	Observed issue
Ettin Decoder 400M (RP)	#1	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.	The US has been working on a new space telescope to study the universe.	Hallucination
	#2	Earthquakes damage all structures, including bridges. Luckily, this kind of collapse is relatively infrequent, especially with modern bridges. Engineers have learned to design bridges in earthquake zones on areas that are much more resistant to movement.	Earthquakes damage all structures, including bridges. Luckily, this kind of collapse is relatively infrequent, especially with modern bridges. Engineers have learned to design bridges in earthquake zones on areas that are much more resistant to movement.	Little or no simplification
	#3	The Hunger Games are an annual event, which the Capitol uses to remind the districts that they are being punished for rebelling. Every year one boy and one girl aged 12 to 18 are chosen from each district by lottery to compete in a battle in which only one person can survive. The chosen teenagers are called "tributes". The whole nation is forced to watch the event on television.	Target CEFR level: A2 Now simplify this one: Original: The Hunger Games are an annual event, which the Capitol uses to remind the districts that they are being punished for rebelling. Every year one boy and one girl aged 12 to 18 are chosen from each district by lottery to compete in a battle in which only one person can survive. The chosen teenagers are called "tributes". The whole nation is forced to watch the event on television.	Repetition
	#4	Many of the major supermarket chains have come under fire with accusations of various unethical acts over the past decade. They've wasted tonnes of food, they've underpaid their suppliers and they've contributed to excessive plastic waste in their packaging, which has had its impact on our environment. But supermarkets and grocers are starting to sit up and take notice. In response to growing consumer backlash against the huge amounts of plastic waste generated by plastic packaging, some of the largest UK supermarkets have signed up to a pact promising to transform packaging and cut plastic wastage.	The supermarket chain Tesco has signed up to a pact promising to transform packaging and cut plastic waste. The supermarket chain has signed up to a pact promising to transform packaging and cut plastic waste. Tesco has signed up to a pact promising to transform packaging and cut plastic waste. Tesco has signed up to a pact promising to transform packaging and cut plastic waste. Tesco has signed up to a pact promising to transform packaging and cut plastic waste. Tesco has signed up to a pact promising to transform packaging and cut plastic waste. Tesco has signed up to a pact promising to transform packaging and cut plastic waste...	Incoherent or low-fluency output
Ettin Decoder 400M (SRP)	#5	Much of today's business is conducted across international borders, and while the majority of the global business community might share the use of English as a common language, the nuances and expectations of business communication might differ greatly from culture to culture. A lack of understanding of the cultural norms and practices of our business acquaintances can result in unfair judgements, misunderstandings and breakdowns in communication. Here are three basic areas of differences in the business etiquette around the world that could help stand you in good stead when you next find yourself working with someone from a different culture.	1. The importance of politeness 2. The importance of respect 3. The importance of etiquette 4. The importance of etiquette in the workplace 5. The importance of etiquette in the home	Hallucination
	#6	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.	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.	Little or no simplification
	#7	The Life of Pi tells the extraordinary story of Pi, a teenage boy from India, who is shipwrecked and finds himself trapped in a lifeboat in the Pacific Ocean with a ferocious tiger. It is the third book by the Canadian author Yann Martel, and was published in 2001. Since then, it has sold over seven million copies worldwide, won several prizes and been translated into 41 languages. Martel is the son of a diplomat and as a child lived in Costa Rica, Canada, France and Mexico.	The Life of Pi is a story about a young boy who is shipwrecked in the Pacific Ocean and finds himself trapped in a lifeboat with a tiger. The tiger is a wild animal and the boy has to survive in the lifeboat. The boy has to survive because he is a vegetarian and he has to feed the tiger with fish and vegetables. The boy has to survive because he is a vegetarian and he has to feed the tiger with fish and vegetables. The boy has to survive because he is a vegetarian and he has to feed the tiger with fish and vegetables...	Incoherent or low-fluency output
	#8	Some bridge collapses are mysteries, and engineers only realise why after they conduct a complete investigation. In some cases, this could happen because inferior-quality material was used in the construction, or because of a defect in a key piece of the bridge. In other cases, the bridge was designed only to support a certain amount of weight and no more.	Some bridges collapse because of poor construction or because of a defect in the bridge itself.	Over simplification and lack of nuance

Table 3: Qualitative analysis of errors produced by Ettin Decoder model simplifying trial data.

Model	Id.	Original text	Simplified text	Observed issue
LLaMA 3 8B (RP)	#9	One night in 2003, Suzanne Collins was flipping between channels on TV. It was at the time of the US invasion of Iraq. The only programmes she could find on TV were 'reality' programmes of young people competing to win a million dollars and news programmes about the war. Suzanne says that the two things start to merge together in a disturbing way and the idea for The Hunger Games came to her.	One night, Suzanne Collins was watching TV. She saw reality shows where people competed to win money and news about the war. She says it was strange how these two things mixed together in a scary way. This made her think of The Hunger Games. Note: I've kept the main ideas and events of the original text, but simplified the language and sentence structure to match the A2 CEFR level.	Rigid, artificial, or robotic output
	#10	I used to do a typical five-day week, but after I came out of my parental leave, I decided that I wanted to spend more time with my children before they start school. After negotiating with my boss, we decided to cut my working week down to a three-day work week. This of course meant a significant cut in my pay too, as I'm paid on a pro-rata basis. I've since noticed, though, that my workload hasn't decreased in the slightest! I'm now doing five days' worth of work in three days, but getting paid much less for it!	I used to work five days a week. After taking time off to be with my children before they start school, I wanted to spend more time with them. I talked to my boss and we agreed to reduce my working hours to three days a week. This means I earn less money because I'm paid based on the number of days I work. But surprisingly, my workload hasn't changed!	Over simplification and lack of nuance
LLaMA 3 8B (SRP)			Here's the rewritten text at CEFR level B1:	
	#11	The Hunger Games are an annual event, which the Capitol uses to remind the districts that they are being punished for rebelling. Every year one boy and one girl aged 12 to 18 are chosen from each district by lottery to compete in a battle in which only one person can survive. The chosen teenagers are called "tributes". The whole nation is forced to watch the event on television.	The Hunger Games happen every year as a reminder of what happens when the districts rebel against the Capitol. As punishment, one boy and one girl between 12 and 18 years old are randomly selected from each district to take part in a deadly competition where only one person can win. These teenagers are called "tributes". The whole country is forced to watch this event on TV. Note: I've used simpler vocabulary and sentence structures to make the text more accessible at CEFR level B1. I've also added some detail, such as mentioning that it's a punishment for rebellion, to help readers understand the context better.	Rigid, artificial, or robotic output
	#12	Wasfia Nazreen first came across a hula hoop as a young girl, when she saw a foreign child who was visiting her native Bangladesh playing with one. Wasfia reluctantly stood by and watched, as in her country it was believed that girls should not play with hula hoops or ride bikes. Now Wasfia is one of the few people in the world to have climbed the Seven Summits, including Everest and Kilimanjaro, and the first to have hula-hooped on each peak. Her reason for doing so: to empower women and girls in a country which discourages them from doing sport.	Here is the rewritten text at CEFR level A2: Wasfia Nazreen saw a hula hoop when she was little. She saw a foreign child playing with it in Bangladesh, where she grew up. In her country, people thought girls shouldn't play with hula hoops or ride bikes. Wasfia just watched. Now, Wasfia is one of the few people who climbed all seven big mountains, including Everest and Kilimanjaro. She's the first person to hula-hoop on each mountain! She did this to help women and girls in her country do sports.	Overloaded output, with unnecessary details

Table 4: Qualitative analysis of errors produced by LLaMA 3 model simplifying trial data.