

# Archaeology at TSAR 2025 Shared Task: Teaching Small Models to do CEFR Simplifications

Rareş-Alexandru Roşcan and Sergiu Nisioi\*

Human Language Technologies Research Center

Faculty of Mathematics and Computer Science

University of Bucharest

roscanrares@gmail.com

sergiu.nisioi@unibuc.ro

## Abstract

Large language models (LLMs) have demonstrated strong performance in text simplification tasks, but their high computational cost and proprietary nature often limit practical use, particularly in educational contexts. In this work, we explore open-source LLMs for CEFR-level text simplification. By reducing the size of the model and the computational requirements, our approach enables greater accessibility and deployment in educational environments. Our results obtain some of the lowest error rates in producing CEFR-compliant texts during TSAR2025, with models of 8 billion and 1 billion parameters. Such approaches have the potential to democratize NLP technologies for real-world applications.

## 1 Introduction

Text simplification (TS) aims to preserve the original meaning while reducing linguistic complexity, making text more accessible to a wider audience (Alva-Manchego et al., 2020). In educational contexts, this is particularly valuable, as learners benefit from materials adapted to their proficiency levels (Li et al., 2025). In this paper, we explore different prompting methods to Large Language Models (LLMs) and provide a comparative analysis between large closed-source systems and smaller open-weights models. While recent LLMs demonstrate strong language understanding capabilities, their high computational requirements and proprietary nature still limit broad educational deployment, particularly in resource-constrained settings (Bai et al., 2024). One-billion parameter models fine-tuned with synthetic data show promising results in providing accessible solutions for educational use.

While prior work has treated CEFR-level TS as a unified task with controllable outputs (Barayan

et al., 2025; Farajidizaji et al., 2024), we propose decomposing this problem into distinct sub-tasks, where each CEFR level represents an independent simplification objective.

We investigate two approaches: (1) level-specific prompts that are very descriptive individually for the specific target, and (2) specialized fine-tuning where models are trained exclusively for individual CEFR levels. Additionally, we use an iterative prompting technique inspired by Minimum Bayes Risk decoding (Heineman et al., 2024), where multiple attempts with slight prompt variations are used to achieve optimal simplification quality.

## 2 Simplification Systems

### 2.1 Prompting

Although prior work used more general prompting strategies (Barayan et al., 2025), we find that highly specific prompts can guide the model more effectively in producing CEFR-aligned simplifications.

Specifically, our prompts make the target CEFR level explicit by including detailed requirements related to vocabulary range, sentence length, and grammatical constructions, as well as by highlighting distinctions between adjacent levels (e.g., A2 vs. B1 and B1 vs B2) while remaining a zero-shot approach. We adopt this method because recent studies have shown that LLMs struggle with reliable CEFR assessment (Barayan et al., 2025; Heineman et al., 2024). The reference to “most common 1500/3000 words” was not based on any external frequency list. Instead, these constraints were communicated in the prompt itself, allowing the LLM to operationalize the notion of common vocabulary internally. The complete prompt templates for each CEFR level are provided in Appendix C and D.

We evaluate our prompting approach across several LLMs to assess their effectiveness in CEFR-aligned text simplification. Specifically, we test GPT-4 Turbo to enable direct comparison with

\*Both corresponding authors.

Barayan et al. (2025), Claude Sonnet 4, Kimi K2, and two LLaMA variants: LLaMA 3.1 8B Instruct and LLaMA 3.2 1B Instruct.

Additionally, we implement an iterative simplification approach where each model is given up to 5 attempts per text to achieve better simplifications, retaining the final iteration as our output.

The strategy is:

1. prompt the model to generate a simplification candidate and evaluate it using the CEFR classifier;
2. the generated text is evaluated for its CEFR level using an ensemble of three BERT-based classifiers (**ModernBERT-base-doc\_en-CEFR**, **ModernBERT-base-doc\_sent\_en-Cefr**, and **ModernBERT-base-reference\_AllLang2-Cefr2**) the same way it is evaluated in the evaluation script;
3. if the simplification matches the target CEFR level, the process stops; otherwise, corrective feedback is provided in the form of The text is TOO COMPLEX. Simplify more aggressively. or The text is TOO SIMPLE. Add more complexity while staying at target level., and the model regenerates a new candidate;
4. the process repeats until the CEFR target is reached or the maximum number of iterations (5) is reached.

**Submission 1** represents the iterative prompting applied to Claude Sonnet 4<sup>1</sup> (claude-4-sonnet-20250514). We use the prompts described in Appendix C and D, with decoding parameters set to  $\text{top}_k = 10$  and  $\text{temperature} = 0.1$ .

**Submission 2** represents the iterative prompting applied to LLaMA 3.1 8B Instruct<sup>2</sup>, using the same prompt templates from Appendix C and D. For decoding, we set  $\text{top}_p = 0.8$  and  $\text{temperature} = 0.1$ .

We evaluate the simplifications using the official TSAR 2025 test dataset and the corresponding evaluation metrics (Alva-Manchego et al., 2025). The test set consists of 100 independent sentences, which are a subset of the UniversalCEFR dataset,

<sup>1</sup><https://www.anthropic.com/claude>

<sup>2</sup><https://huggingface.co/meta-LLaMA/LLaMA-3.1-8B-Instruct>

each paired with two human-authored simplifications: one at the A2 level and one at the B1 level.

## 2.2 Synthetic Data

UniversalCEFR dataset (Imperial et al., 2025) is a corpus annotated with CEFR levels without parallel versions in simplified form. To address this issue, we use LLaMA 3.1 8b to simplify the texts at the target CEFR levels by prompting the model with the iterative method presented earlier.

We select all texts annotated at B2 level or above, both at the document and sentence level. Sentence-level texts are used directly, while document-level texts are parsed into smaller texts. In total, we obtain 12,000 distinct texts, each paired with two simplified versions: one at A2 level and one at B1 level.

## 2.3 Finetuning Small models

We use the synthetic dataset described in subsection 2.2 to finetune LLaMA 3.2 1B Instruct for CEFR-based text simplification. We experiment with two finetuning setups, in the first approach, we train a single model on the entire dataset, such that it learns to generate both A2 and B1 level simplifications. In the second approach, we train two separate models, one specialized for A2 simplification and the other for B1. We perform fine-tuning using the Axolotl framework<sup>3</sup>, which provides training pipelines for instruction-tuned LLaMA models.

Our training configuration uses a learning rate of  $5e-5$  with cosine learning rate scheduling and the AdamW optimizer with 8-bit quantization. We set the micro-batch size to 1 with gradient accumulation steps of 32, resulting in an effective batch size of 32. The models are trained for 2 epochs with a maximum sequence length of 4096 tokens, using sample packing to improve training efficiency. We apply a weight decay of 0.001 and use automatic mixed precision (bf16) when available. The validation set comprises 10% of the training data, and we evaluate model performance every 50 steps using perplexity as the primary metric. Early stopping is not employed, and we save model checkpoints once per epoch.

**Submission 3** represents the iterative prompting to LLaMA 3.2 1B Instruct<sup>4</sup>, fine-tuned using the configuration mentioned in subsection 2.3. We use

<sup>3</sup><https://github.com/OpenAccess-AI-Collective/axolotl>

<sup>4</sup><https://huggingface.co/meta-LLaMA/LLaMA-3.2-1B-Instruct>

the same iterative prompting process and prompt templates as in previous submissions.

### 3 Results and Discussion

Table 1 summarizes the performance of all submissions across the evaluation metrics. Overall, the results indicate that Claude Sonnet 4 achieves the best performance, with the lowest RMSE (0.122) and the highest automatic ranking score (1.0). This highlights its strong ability to reliably adjust lexical and syntactic complexity while preserving meaning, this being the reason why we chose it for synthetic data.

While Kimi K2 achieves strong results across automatic metrics, qualitative inspection suggests that its simplifications often lack fluency and cohesion. In particular, the model tends to produce excessively fragmented and repetitive sentences, which explains the gap between metric-based performance and actual readability. This pattern also raises the possibility that Kimi K2 may be exploiting weaknesses in current evaluation metrics, producing outputs that score well numerically. A detailed comparison with human references and outputs from Claude and LLaMA 3.1 is provided in Appendix 2.

In contrast, the results for LLaMA 3.1 8B Instruct illustrate a divergence between automatic evaluation and human judgment. According to Table 1, this model ranks below Kimi K2, with a higher RMSE and slightly lower automatic scores. However, our manual analysis found the outputs of LLaMA 3.1 8B to be more coherent, fluent, and contextually appropriate. This discrepancy underscores the fact that automatic metrics alone do not fully capture aspects such as fluency and cohesion, which are essential in CEFR-based text simplification. Consequently, we regard LLaMA 3.1 8B as overall stronger in practical terms, despite its lower position in the automatic ranking.

When examining the outputs of LLaMA 3.2 1B Instruct without fine-tuning, two issues become evident. First, the model often fails to follow all prompt instructions, even when these are stated multiple times. For example, it frequently returns additional comments or explanations such as “Here is the simplified text...” or even detailed notes about the transformation process, despite being explicitly instructed to output only the simplified text. Second, the base model tends to produce excessively short simplifications, sometimes reducing an entire

paragraph to one or two sentences. This extreme shortening is caused partly by the iterative prompting setup: since the model has no prior knowledge of how a CEFR-target simplification should look (zero-shot), the repeated corrections encourage it to strip down content aggressively rather than restructure it appropriately.

Moreover, the extra notes and process explanations negatively bias the iterative pipeline: because the CEFR classifier evaluates the entire response, the added commentary shifts the predicted level upwards or downwards in unintended ways. For instance, in one case the model returned a step-by-step justification of lexical changes (“I removed ‘common’ from ‘Common dreams’, changed...”), which was treated as part of the text by the classifier and distorted the CEFR label. Examples of both the over-shortening and adherence failures are documented in Appendix 3, demonstrating how iterative prompting without prior fine-tuning leads to unreliable outputs and noisy automatic evaluations.

By contrast, the fine-tuned 1B model overcomes these limitations: it aligns much more closely with the intended structure of CEFR simplifications, avoids adding unwanted commentary, and produces outputs that are considerably richer in content while still controlled in complexity. These improvements are particularly noteworthy given that the fine-tuning relied on synthetic data generated with LLaMA 3.1 8B, itself a relatively modest model, and that the resulting dataset was not manually cleaned or verified. This suggests that the gains observed here represent only a lower bound of what could be achieved; with higher-quality, carefully curated, or state-of-the-art synthetic simplification data, fine-tuning small models would likely provide even greater improvements.

For completeness, we also conducted a small-scale fine-tuning experiment on LLaMA 3.1 8B, using approximately 1000 synthetic simplifications generated with Claude Sonnet 4 and the same fine-tune configuration as the one used for LLaMa 3.2 1B. While the dataset was limited in size and not manually verified, it still allowed us to observe meaningful improvements. Due to the small amount of data, we did not apply iterative prompting during inference and instead evaluated the model in a single-try setting. This setup affected the results in two ways: (1) RMSE scores were higher compared to the iterative experiments, since the model had fewer opportunities to adjust toward the exact CEFR level, and (2) the Mean-

Model	RMSE	MeaningBERT (orig)	MeaningBERT (ref)	Avg Score	Auto Rank
<b>Claude Sonnet 4<sup>1</sup></b>	0.122	0.779	0.804	0.338	1.000
LLaMA 3.1 8B finetune non-iterative	0.324	0.807	0.822	0.195	1.710
<b>LLaMA 3.1 8B<sup>2</sup></b>	0.255	0.781	0.789	0.135	2.000
Kimi K2	0.173	0.726	0.771	0.099	2.180
GPT-4 Turbo	0.274	0.730	0.776	0.000	2.670
<b>LLaMA 3.2 1B individual finetune<sup>3</sup></b>	0.212	0.706	0.731	-0.104	3.180
LLaMA 3.2 1B finetune	0.300	0.721	0.738	-0.167	3.490
LLaMA 3.1 8B non-iterative	0.561	0.794	0.795	-0.192	3.610
LLaMA 3.2 1B	0.663	0.574	0.588	-1.285	9.000

Table 1: AUTORANK results for all submitted runs under custom weighting. <sup>1,2,3</sup> mark our official submissions. “Individual” refers to individually fine-tuned models while “non-iterative” refers to single try simplifications. If not mentioned, all models are iterative/nonfinetuned. The other models are included for comparison.

ingBERT scores improved, consistent with our hypothesis that the first simplification attempt usually yields the highest semantic similarity. Even under these constrained conditions, the one-try RMSE remained competitive, showing that even with a very limited number of training examples, fine-tuning substantially enhanced the model’s ability to perform CEFR simplification.

It is also important to note that our experiments were primarily optimized towards minimizing RMSE, as this metric directly reflects the alignment of the outputs with the target CEFR level. Consequently, aspects related to meaning preservation were not prioritized to the same extent. This explains why some systems, despite achieving lower RMSE, occasionally sacrificed fidelity or omitted information from the original text. A more balanced approach that explicitly weights both accuracy of CEFR control and semantic consistency could represent an important direction for future work.

Furthermore, our focus on minimizing RMSE exposed some of the limitations and biases in the current evaluation metrics. For example, Kimi K2 achieves relatively strong scores despite producing outputs that are qualitatively weak, highlighting that automatic metrics can sometimes reward undesirable behavior. This suggests that more robust, state-of-the-art evaluation methods are necessary to obtain a fair ranking of systems. Under such an evaluation framework, we hypothesize that an approach built on well-balanced Minimum Bayes Risk decoding could result in models that are both reliable for CEFR control and practical enough to be deployed.

In addition, as shown in Table 1, fine-tuning two

separate LLaMA 3.2 1B models individually for A2 and B1 simplification yields lower slightly RMSE compared to a single model trained on both levels. This indicates that task specialization, even at very small model scales, can provide improvements in controlling output complexity.

#### 4 Ethical and Deployment Considerations

Our methods enable practical deployment of CEFR-aware simplification, but operational use raises concrete risks. Synthetic training data and automatic simplifiers may introduce factual inaccuracies, level misclassification, or unequal behaviour across learner groups. Therefore, any real-world deployment should include human validation, continuous monitoring for meaning preservation and fairness, clear documentation of model limitations and licensing, and user-facing transparency about automated processing.

Furthermore, the proposed approach contributes to a more sustainable deployment paradigm: instead of relying on repeated inference from large proprietary LLMs, it enables the generation of synthetic data to train smaller, open models. This reduces computational and financial costs while improving accessibility for educational institutions with limited resources.

Another advantage of our method lies in its applicability to low-resource settings, where high-quality labeled simplification data are scarce. Synthetic data generation through LLMs can bootstrap new CEFR-aligned datasets for such languages.

Future work could incorporate human-in-the-loop feedback from language teachers to refine level control and to assess pedagogical usefulness beyond automatic metrics.

## 5 Conclusions

This work demonstrates several key findings for CEFR-based text simplification using smaller, open-source language models.

Our iterative prompting approach provides a computationally inexpensive method to significantly improve simplification quality. By allowing models up to 5 attempts per text, we achieved substantial RMSE improvements across all tested models. This technique offers a practical way to enhance performance without additional training or model scaling, making it particularly valuable for resource-constrained educational environments.

Highly structured, CEFR-specific prompts improve model understanding of proficiency level requirements. Our detailed prompts, which explicitly describe vocabulary constraints, sentence complexity, and grammatical structures for A2 and B1 levels, led to more accurate simplifications compared to generic approaches. This finding suggests that explicit linguistic guidance can compensate for models' limited understanding of CEFR frameworks.

Training separate models for each CEFR level yields better results than training a single model to handle multiple target levels. This specialization allows models to develop more focused representations of the linguistic constraints and stylistic requirements specific to each proficiency level, resulting in more appropriate simplifications.

Our fine-tuning experiments further serve as a strong proof of concept. We observed that even when training with synthetic data generated by LLMs (which were not manually verified and may inherently contain too much noise) substantial improvements can still be achieved. This effect was visible both when using large amounts of automatically generated samples (from LLaMA 3.1 8B) and under very constrained conditions with only 1000 examples (from Claude Sonnet 4 for LLaMA 3.1 8B). These results highlight the potential of fine-tuning: with higher-quality and carefully curated datasets, the gains observed here would likely be amplified, suggesting a clear path forward for building robust simplification systems on top of smaller open models for educational use.

Our code is available at <https://github.com/roscaanraes/emnlp-2025-tsar>.

## 6 Lay Summary

This study explores how large language models (LLMs) can be guided to produce simplified texts that match the Common European Framework of Reference for Languages (CEFR) levels. Our aim is to make reading materials easier to understand by automatically adjusting text difficulty to the reader's language level.

Many previous studies have used general instructions or relied on large datasets labeled by hand. Our approach experiments with CEFR-aware prompting, which gives the model more specific instructions to generate simplified examples that match the language characteristics of each level. These examples can then be used to train or evaluate smaller, open-access models, reducing the need for expensive proprietary LLMs.

We are also studying how the level of detail in these instructions affects the model's ability to produce appropriate simplifications. Early results suggest that more specific prompts can improve control over vocabulary and sentence structure, leading to texts that are easier to read and score better in automatic evaluations.

In addition, we observe that this method enables a more sustainable way of developing simplification systems. Instead of relying on repeated queries to large commercial models, synthetic data produced once can be reused to fine-tune compact open models. This reduces computational costs, encourages reproducibility, and supports educational organizations that may not have access to high-end infrastructure.

This work is still in progress. While our findings are promising, more research is needed before such methods can be turned into practical tools. Ultimately, we hope this research will help teachers and educational institutions create more accessible, level-adjusted learning materials and support the development of CEFR-aligned datasets for languages with limited resources.

## Acknowledgments

This work was supported by the Romanian National Research Council (CNCS) through the Executive Agency for Higher Education, Research, Development and Innovation Funding (UEFISCDI) under grant PN-IV-P2-2.1-TE-2023-2007 InstRead.

## References

- 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.
- Guangji Bai, Zheng Chai, Chen Ling, Shiyu Wang, Jiaying Lu, Nan Zhang, Tingwei Shi, Ziyang Yu, Mengdan Zhu, Yifei Zhang, Xinyuan Song, Carl Yang, Yue Cheng, and Liang Zhao. 2024. [Beyond efficiency: A systematic survey of resource-efficient large language models](#). *Preprint*, arXiv:2401.00625.
- 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.
- 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, Torino, Italia. ELRA and ICCL.
- David Heineman, Yao Dou, and Wei Xu. 2024. [Improving minimum bayes risk decoding with multi-prompt](#). *Preprint*, arXiv:2407.15343.
- Joseph Marvin Imperial, Abdullah Barayan, Regina Stodden, Rodrigo Wilkens, Ricardo Muñoz Sánchez, Lingyun Gao, Melissa Torgbi, Dawn Knight, Gail Forey, Reka R. Jablonkai, Ekaterina Kochmar, Robert Reynolds, Eugénio Ribeiro, Horacio Saggion, Elena Volodina, Sowmya Vajjala, Thomas François, Fernando Alva-Manchego, and Harish Tayyar Madabushi. 2025. [UniversalCEFR: Enabling Open Multilingual Research on Language Proficiency Assessment](#). *arXiv preprint arXiv:2506.01419*.
- Guanlin Li, Yuki Arase, and Noel Crespi. 2025. [Aligning sentence simplification with ESL learner’s proficiency for language acquisition](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 492–507, Albuquerque, New Mexico. Association for Computational Linguistics.

## A Things we Thought Would Work, but didn’t

In addition to our successful experiments, we report several approaches that produced suboptimal results, providing valuable insights for future research in CEFR-based text simplification.

### A.1 Extended Synthetic Training Data

We initially hypothesized that including intermediate simplification attempts (not just the final outputs) would provide richer training signals. However, expanding our training dataset from 12,000 to 48,000 examples by including all intermediate simplifications introduced excessive noise and degraded model performance. This indicates that quality of training examples is more critical than quantity in this context and that it might be difficult to finetune for iterative simplifications.

### A.2 Alternative Model Architectures

Experiments with Qwen 32B yielded considerably weaker results compared to LLaMA 3.1 8B. Both the base version of Qwen and its fine-tuned variant underperformed relative to LLaMA 3.1 8B across metrics such as RMSE and MeaningBERT, indicating that model size alone does not dictate performance.

In addition, we also attempted fine-tuning on several other open-source models, including EXAONE-4.0-1.2B<sup>5</sup>, OLMo-2-0425-1B-Instruct<sup>6</sup>, and LFM2-1.2B<sup>7</sup>. However, none of these models produced competitive results in our setup.

### A.3 Sub-1B Parameter Models

We experimented with several ultra-lightweight models, including Gemma variants and LFM2-350M<sup>8</sup> systems with fewer than 1 billion parameters. While these models were able to return outputs in a one-shot setup, their results were significantly weaker than those of LLaMA 3.2 1B, both in terms of RMSE and semantic preservation. The shortcomings became even more apparent under iterative prompting, where they showed the same type of failures as the unfine-tuned LLaMA 3.2 1B, including hallucinations and unstable outputs. Given that

<sup>5</sup><https://huggingface.co/LGAI-EXAONE/EXAONE-4.0-1.2B>

<sup>6</sup><https://huggingface.co/allenai/OLMo-2-0425-1B-Instruct>

<sup>7</sup><https://huggingface.co/LiquidAI/LFM2-1.2B>

<sup>8</sup><https://huggingface.co/LiquidAI/LFM2-350M>

even their one-try simplifications were of very low quality, we conclude that such sub-1B models are not a viable choice for CEFR simplification in our setup, especially when compared to the much stronger performance achieved by LLaMA 3-series models.

#### A.4 Cross-lingual Performance Variations

Models such as Kimi K2 and Qwen, which were likely trained on larger Chinese language corpora, showed different processing patterns and performance characteristics compared to models with more balanced multilingual training. This highlights the importance of pre-training data composition for downstream task performance.

#### A.5 LLaMA Models: Exceptional Performance-to-Cost Ratio

Remarkably, our experiments reveal that LLaMA variants, particularly LLaMA 3.1 8B and even the compact 3.2 1B model, achieve competitive or superior performance compared to much larger proprietary models like GPT-4 Turbo in CEFR simplification tasks. This finding has significant implications:

- **Cost-effectiveness:** Smaller LLaMA models provide accessible alternatives for educational institutions with limited computational resources
- **Synthetic data generation:** Larger LLaMA variants (e.g., LLaMA 4 17B, LLaMA 3.3 70B) could serve as ideal candidates for generating high-quality synthetic training data for even smaller specialized models
- **Deployment feasibility:** The strong performance of 1B parameter models opens possibilities for on-device deployment in educational applications

#### A.6 CEFR Assessment Finetune

We also experimented with joint fine-tuning approaches on LLaMA 3.1 8B that combined CEFR assessment with simplification. Our motivation for this experiment was the observation that models generally lack the capacity to reliably perform CEFR assessment on their own, which led us to hypothesize that an explicit assessment signal could help guide the simplification process. Specifically, we tested two strategies: (1) sequential fine-tuning, where we first trained the model for CEFR text

classification and then adapted it for simplification, and (2) multi-task fine-tuning, where we included both simplification data and CEFR assessment data in the same training stage.

For the assessment task, we relied on the same UniversalCEFR dataset used in our synthetic data experiments, but instead of parsing document-level texts into smaller segments, we used the full documents together with the sentence-level samples, resulting in approximately 12,000 texts. For the simplification task, we used the same synthetic simplifications generated by Claude Sonnet 4. However, neither approach produced improvements over standard fine-tuning. These results suggest that directly combining CEFR assessment and simplification is not straightforward, and that more advanced methods such as curriculum learning or more carefully balanced multi-task objectives may be required.

## B Models and Synthetic Dataset

All fine-tuned LLaMA models used in this work are publicly available on Hugging Face:

- **LLaMA 3.1 8B fine-tuned for A2 and B1:**  
<https://huggingface.co/roscanraes/llama-3.1-8b-finetuned-a2b1>
- **LLaMA 3.2 1B fine-tuned for A2:**  
<https://huggingface.co/roscanraes/llama-3.2-1b-finetuned-a2>
- **LLaMA 3.2 1B fine-tuned for B1:**  
<https://huggingface.co/roscanraes/llama-3.2-1b-finetuned-b1>

The synthetic dataset used for training and evaluation can be downloaded from:

- **CEFR-Simplifications dataset:**  
<https://huggingface.co/datasets/roscanraes/CEFR-Simplifications>

## **C A2 Prompt**

### **Role**

You are a language teacher simplifying texts to A2 CEFR level.

### **Objective**

Transform this text to A2 level while preserving all original meaning and information. {base\_feedback}

### **A2 Language Requirements**

- Vocabulary: Most common 1500 English words only
- Sentences: 8–12 words, one clear idea per sentence
- Grammar: Simple present/past, basic future (will), basic modals (can/must/should)
- Connectors: and, but, because, so, when, if, then
- Style: Personal, concrete, everyday language

### **Strict Level Control**

- Above A1: Include personal experiences, feelings, plans, time references
- Below B1: No present perfect, passive voice, or complex connectors (however, although, despite)
- Below B1: No abstract concepts without concrete explanation

### **Transformation Process**

1. Identify all key information and meaning
2. Break complex sentences into simple A2 structures
3. Replace advanced vocabulary with A2 equivalents
4. Convert complex grammar to simple A2 patterns
5. Verify all original meaning is preserved

### **Critical**

Do not omit, summarize, or change any information. Only change **how** it is expressed.

### **Instruction**

Return only the simplified text. Do not include any other comments, notes, or additional information.

### **Text to Simplify**

## **D B1 Prompt**

### **Role**

You are an expert CEFR B1 text simplification specialist with deep understanding of automatic language assessment systems.

### **Objective**

Transform this text to precise B1 level while preserving all original meaning and information. {base\_feedback}



## **B1 Language Requirements**

- Vocabulary: 2000–3000 most common English words, avoid academic/formal terms
- Sentences: 15–22 words, can connect 2 related ideas with clear logic
- Grammar: Present perfect (have/has done), simple passive (is/was done), basic conditionals (if...will/would), modals (should, might, could, would)
- Connectors: however, although, while, since, unless, because, so that, even though
- Style: Clear intermediate language that shows reasoning and personal opinions

## **Strict Level Control**

- Above A2: Include abstract concepts with simple explanation, cause-effect relationships, personal opinions with basic justification, intermediate grammar patterns
- Below B2: No academic/formal vocabulary (facilitate → help, demonstrate → show, utilize → use)
- Below B2: No complex conditional structures, no sophisticated argumentation, no specialized terminology without explanation
- Precise B1 Target: Intermediate complexity using everyday vocabulary — never oversimplify to A2, never undersimplify leaving B2+ elements

## **Critical B1 Differentiators**

- From A2: Can handle abstract ideas but explains them simply using common words
- From B2: Uses everyday vocabulary even for complex concepts, avoids formal/academic tone
- B1 Signature: Connects ideas logically but with simple language patterns

## **Transformation Process**

1. Identify all key information and meaning
2. Scan for B2+ vocabulary and replace with B1 common equivalents
3. Convert complex sentences to B1 structures (maximum 2 clauses per sentence)
4. Add simple explanations for any remaining complex concepts
5. Include 2–3 B1 grammar markers per paragraph naturally
6. Verify consistent B1 complexity throughout — no A2 oversimplification, no B2+ elements remaining

## **Critical**

Do not omit, summarize, or change any information. Only change **how** it is expressed to match B1 patterns that automatic CEFR classifiers consistently recognize as B1 level.

## **Instruction**

Return only the simplified text. Do not include any other comments, notes, or additional information.

## **Text to Simplify**

Source	Text
Original	Your dreams might show you all kinds of insights into things that you didn't realise you were thinking about. Common dreams like being able to fly or falling, your teeth falling out or having no clothes on in a public place probably mean something similar in most people. But the key to understanding exactly what they mean to you is to connect them to the events and feelings in your daily life.
Kimi K2	Your dreams can show you things. They show what you think about. Many people have the same dreams. They dream they can fly. They dream they fall. They dream their teeth fall out. They dream they have no clothes in a public place. These dreams mean the same thing for most people. But you must connect the dream to your life. You must connect it to your feelings. Then you will know what it means to you.
Claude Sonnet 4	Your dreams can show you new ideas about things. You did not know you were thinking about these things before. Many people have the same dreams. Some people dream they can fly. Some people dream they are falling. Some people dream their teeth fall out. Some people dream they have no clothes on when other people can see them. These dreams probably mean similar things for most people. But you must think about your own life to understand your dreams. You must think about what happens in your day. You must think about how you feel every day. Then you can understand what your dreams mean for you.
LLaMA 3.1 8B	Your dreams can show you things you didn't think about before. Many people have dreams like flying or falling. These dreams are probably the same for everyone. To understand your dreams, you must think about what you do and feel every day.
Human Reference	People have different dreams when they sleep. Some common dreams are flying, falling, losing teeth, or being outside with no clothes. These dreams often mean the same thing for many people. To know what your dream means, think about what is happening in your life and how you feel.

Table 2: Example simplifications produced by different models compared to the human reference.

<b>Example 1: Oversimplification (very short output)</b>	
<b>Original</b>	Social media, magazines and shop windows bombard people daily with things to buy, and British consumers are buying more clothes and shoes than ever before. Online shopping means it is easy for customers to buy without thinking, while major brands offer such cheap clothes that they can be treated like disposable items – worn two or three times and then thrown away.
<b>Reference (human)</b>	Every day, people see many clothes and shoes on social media, in magazines, and in shops. In the UK, people are buying more than before. Online shopping makes it easy to buy without thinking. Big brands sell clothes that are very cheap. Some people wear them only a few times and then throw them away.
<b>1B No Fine-tune</b>	People buy things on line. They do not want these things.
<b>1B Fine-tuned</b>	People see many ads on social media and in shops every day. They want to buy things. British people buy more clothes and shoes now than ever before. Online shopping is easy. You don't need to think about it. Major brands sell cheap clothes. These clothes can be used once or twice and then thrown away.
<b>Example 2: Added explanations despite prompt restrictions</b>	
<b>Original</b>	A large number of bridge accidents occur during the construction of the bridge itself. These accidents are often due to an error made by the engineers, such as a miscalculation. The bridge collapses under its own weight, and this can be deadly for the workers on it at the time.
<b>Reference (human)</b>	Many accidents with bridges happen while they are being built. These accidents often happen because engineers make mistakes, like wrong numbers. The bridge can fall down from its own weight. This is very dangerous for the workers on the bridge.
<b>1B No Fine-tune</b>	Here is the simplified text: "Many people get hurt when they drive over bridges. Sometimes there is a mistake in how the bridge is built, like making a wrong calculation. This makes the bridge fall down and it can be very bad for those who work on it."
<b>1B Fine-tuned</b>	Many people get hurt in bridge building accidents. This happens because of a mistake made by the builders. They make a wrong calculation, and the bridge falls down. It can be very dangerous for the workers who are working on it.

Table 3: Outputs from LLaMA 3.2 1B base vs. fine-tuned models compared against original and human reference simplifications. Example 1 illustrates oversimplification by the base model, while Example 2 shows unnecessary explanations being added despite task instructions.