GRIPF at TSAR 2025 Shared Task: Towards controlled CEFR level simplification with the help of inter-model interactions

David Alfter
GRIDH
Literature, History of Ideas and Religion
University of Gothenburg, Sweden
david.alfter@gu.se

Sebastian Gombert
Educational Technologies
Information Centre for Education
DIPF, Germany
s.gombert@dipf.de

Abstract

In this contribution to the CEFR level simplification TSAR 2025 Shared Task, we propose two systems, *EZ-SCALAR* and *SAGA*, that implement two differing approaches to prompting LLMs for proficiency-adapted simplification. Our results place us in the middle of the participating teams, and reveal that using external lexical resources to guide simplification improves overall results.

1 Introduction

This paper presents the details of the GRIPF team in the TSAR 2025 Shared Task on CEFR level simplification (Alva-Manchego et al., 2025). The workshop website describes the task as follows: "The task targets English-language paragraphs written at upper-intermediate or advanced levels and requires participants to simplify them according to a specified target readability level, defined using the Common European Framework of Reference for Languages (CEFR). Specifically, participants will be asked to simplify texts originally at B2 level or above to target levels of A1, A2, or B1" (TSAR Workshop).[1]

2 Related Work

As Bahrainian et al. (2024) put it, "[t]ext simplification is the process of rewriting a piece of text using simpler vocabulary and grammatical structure in order to make the text more accessible and understandable for a larger audience".

Text simplification has evolved from rule-based approaches that rely on predefined linguistic transformations to neural methods that learn simplification patterns from data. Early work focused on lexical substitution and syntactic restructuring using hand-crafted rules, while recent advances have leveraged transformer-based models for end-to-end simplification. Zetsu et al. (2024) demonstrate that lexically constrained decoding with edit operations can effectively guide neural simplification models, addressing the challenge of loose constraints in previous approaches.

Controllable text generation has emerged as a particularly relevant area, where models are guided to produce outputs meeting specific criteria such as reading level or target audience. Cripwell et al. (2023) propose document-level planning approaches that decompose simplification into discrete operations (copy, rephrase, split, or delete), showing that structured planning can guide generation more effectively than end-to-end approaches. Recent work by Maddela and Alva-Manchego (2025) addresses the challenge of evaluating document-level simplification, proposing methods to adapt sentence-level metrics for longer texts. Within this landscape, CEFR-level simplification represents a specialized challenge, as it requires not only reducing complexity but doing so in alignment with established language proficiency standards. Ryan et al. (2023) highlight the particular challenges of multilingual simplification, while Horiguchi et al. (2025) extend this work by developing parallel corpora for the simplification of medical texts in nine languages, demonstrating that multilingual training can improve the performance of simplification.

Ensemble methods in natural language processing have demonstrated consistent improvements over single-model approaches across various tasks. The principle of combining multiple models to leverage their complementary strengths has been successfully applied to machine translation, text generation, and classification tasks. Recent work by Valiev and Tutubalina (2024) shows that in-prompt ensemble methods, where multiple model predictions are integrated as separate expert solutions with trust scores, can achieve strong per-

[1] While the webpage specifies the target level A1, the actual task only targets A2 and B1.

Proceedings of the Fourth Workshop on Text Simplification, Accessibility and Readability (TSAR 2025), pages 137–148
November 4-9, 2025 ©2025 Association for Computational Linguistics

formance in specialized domains. Parfenova and Pfeffer (2025) demonstrate that smaller ensemble models with refined outputs can consistently outperform individual models and even large-scale LLMs, suggesting that ensemble approaches may be more effective than relying solely on large standalone models. Niess and Kern (2025) extend ensemble methods to watermarking applications, showing that multi-feature ensemble approaches achieve 98% detection rates and maintain robustness against paraphrasing attacks.

Our work extends this concept to controllable simplification, where inter-model critique and refinement can potentially address the limitations of individual models. This approach aligns with recent trends in self-correction and iterative improvement in language models, where systems refine their output through multiple generation cycles. Pan et al. (2024) provide a comprehensive survey of automated correction strategies, categorizing approaches into training-time, generation-time, and post-hoc methods, while Ferraz et al. (2024) demonstrate that decompose-critique-refine pipelines can significantly improve constraint following in language models. However, recent critical surveys suggest that the effectiveness of self-correction varies significantly across tasks and model types (Kamoi et al., 2024), highlighting the importance of ensemble approaches that can leverage external critique mechanisms rather than relying solely on self-evaluation.

3 Methodology

We use two similar approaches to tackle the problem of lexical simplification. The first system is called EZ-SCALAR (Ensemble Zero-shot Simplify, Criticize, Adapt with Lexical Assessment, and Referee). The second system is called SAGA (Self-Assessed Guided Adaptation). Both systems estimate complex vocabulary, either using external resources or by prompting the model. Both models adjust the simplification based on the identified complex vocabulary.

3.1 EZ-SCALAR

EZ-SCALAR uses two state-of-the-art LLMs: GPT-5 and Claude Opus 4.1. In the first step, both models are prompted to simplify the text to the target level. In the second step, each model receives the simplified text of the other model and is asked to critique the simplification given the target level. The first stage generates independent simplifications, allowing each model to apply its learned patterns without bias from the other. The critique stage is crucial for identifying potential issues that individual models might miss, as different architectures often exhibit distinct biases. In the third step, each model receives the other model's critiqueand is asked to refine its text based on the critique; the refinement stage incorporates these insights, potentially producing outputs superior to either model alone. In the last step, each model receives the original text, the two simplifications, and the target level, and is asked to choose the text that best fits the target level in terms of **meaning preservation**, **readability** and **clarity**. In case the models disagree, a third model – Llama-3.2-3B-Instruct – is used as judge.

3.2 EZ-SCALAR Lex

As an extension of the base system, we also experiment with using external lexical resources. The system is identical to EZ-SCALAR, but between the second and third step, a lexical assessment module is implemented. Figure 1 in the Appendix illustrates both architectures. This module uses the EFLLex vocabulary list (Dürlich and François, 2018) to estimate the complexity of vocabulary and addresses a key limitation of purely neural approaches: the lack of explicit control over vocabulary complexity. By integrating the EFLLex vocabulary list, the system bridges rule-based and neural methods, providing concrete guidance about problematic words. Since the EFLLex list contains words with their part-of-speech and frequency distributions over CEFR levels, we use the first-occurrence approach to map each (word, pos) tuple to its target level. This means that each word is linked to a distribution of frequencies over levels (how often was this word used at A1, A2, etc.), and the level at which the frequency first becomes non-zero is taken as target level. Since this method of level assignment is brittle when it comes to low frequencies, all levels above A1 are shifted downwards one level. The decision to shift CEFR levels downward reflects practical considerations about the reliability of frequency-based level assignments, acknowledging that conservative estimates are preferable to overly optimistic ones in educational contexts. The module uses stanza to lemmatize and pos-tag. Since the pos-tags are different, we map the lexical tags *nouns*, *verbs*, *adjectives* and *adverbs* to a common tag set (the tag

set used by EFLLex).[2] The module takes as input the simplified text and target level and identifies words which are above the target level. These words are then added to the prompt for the third step, along with a prompt to pay special attention to these words.

3.3 SAGA

SAGA uses two components, which we refer to as the *proposer* and the *reviewer*. Given a text and a desired CEFR level, the proposer generates an adjusted version of the text that supposedly satisfies that CEFR level. The reviewer then classifies whether the proposed version of the text actually meets the required level. In case the *reviewer* predicts that the text is indeed satisfying the level, we accept. Otherwise, the *proposer* needs to regenerate a text based on the previously generated version, which is then again judged by the *reviewer*. This process is repeated until an agreement is reached or a maximum number of iterations is reached. Figure 2 in the Appendix illustrates the SAGA architecture.

For our submission, the *proposer* is implemented with *GPT-4o* via the conversational API. Given, a custom system prompt (see appendix), the model is first prompted to identify words and phrases within the given input text that are not appropriate for the desired CEFR level and also to provide more appropriate alternatives along with the identified words and phrases. In the second step, the model is then prompted to reformulate the input text by changing the respective phrases and using the proposed alternatives instead.

The resulting text is then given to the *reviewer*, which, in our case is implemented in the form of a *ModernBERT-large* (Warner et al., 2025) classification model fine-tuned on a corpus of English texts labelled for their CEFR level[3] (1x5 cross-validation resulted in a weighted F1 of 0.689). If the *reviewer* detects a different CEFR level than expected, the *proposer* is prompted to identify additional words and phrases that could have brought the reviewer to this judgement, again with suitable alternatives. Following this, it is again prompted to reformulate the text accordingly. For the shared task submission, we set the maximum iterations to 5.

[2]We surmise that marking interjections and other closed-class part-of-speech may be unnecessary and introduce unnecessary noise in the process.

[3]https://www.kaggle.com/datasets/amontgomerie/cefr-levelled-english-texts

4 Evaluation

4.1 Pre-Evaluation

For pre-evaluating our systems, we used the provided trial data set and the official evaluation script. The evaluation script outputs different measures, organized into three overarching parts, each with sub-measures:

1. CEFR Compliance: how well does the text adhere to the specified CEFR level; this compliance is checked by running the simplified text through a model trained on sentence-label CEFR data
    - Weighted F1: This is a combination of precision and recall, averaged by class and weighted by the number of actual occurrences of each class in the dataset (*weighted_f1* in Table 1). It was dropped in the final evaluation because it does not capture the severity of misclassification.
    - Adjacent accuracy: This measure is similar to accuracy, but it counts a prediction as correct if it corresponds either directly to the gold label or is either one above or below the gold label (e.g., if the gold label is B1, and the prediction is A2, it is counted as correct; *adj_accuracy* in Table 1). It was dropped in the final evaluation because it is less informative than RMSE.
    - Root Mean Squared Error (RMSE): RMSE is the quadratic mean of the differences between the observed values and predicted ones (*rmse* in Table 1)
2. Meaning Preservation: how well does the simplification preserve the meaning of the original unsimplified text
    - MeaningBERT-Orig: measures semantic similarity between the original text and the simplified version using a model specifically trained on human annotations for meaning preservation in simplification
    - BERTScore-Orig: measures semantic similarity between the original text and the simplified version. However, as this model tends to "overestimate similarity when there is lexical overlap with no true meaning preservation" (Alva-Manchego

Metric	Sub-Metric	EZ-SCALAR	EZ-SCALAR Lex	SAGA
CEFR Compliance	weighted_f1 ↑	0.533	0.567	0.530
	adj_accuracy ↑	0.975	0.975	0.975
	rmse ↓	0.758	0.742	0.742
Meaning Preservation	MeaningBERT-Orig ↑	0.852	0.860	0.790
	BERTScore-Orig ↑	0.948	0.951	0.951
Similarity to References	MeaningBERT-Ref ↑	0.803	0.794	0.738
	BERTScore-Ref ↑	0.926	0.925	0.920

Table 1: Results on trial data. ↑ means that a higher score (maximum 1) is better. ↓ means that a lower score is better (minimum 0)

System	RMSE ↓	M-BERT-Orig ↑	M-BERT-Ref ↑	AvgScore ↑	AUTORANK ↓
EZ-SCALAR	0.721	0.856	**0.824**	0.060	8.270
EZ-SCALAR Lex	**0.689**	**0.857**	0.820	**0.070**	**8.130**
SAGA	0.831	0.827	0.796	-0.140	10.780

Table 2: Results on test data. Best results per test measure in bold.

et al., 2025), it was dropped for the final evaluation

3. Similarity to References:
   - MeaningBERT-Ref: same as above, except between a human reference annotation and the system simplification
   - BERTScore-Ref: same as above

Our results are summarized in Table 1. However, as the number of data points in the trial data set was limited ($n = 40$), the results are merely hints at what methods might work better.

As can be gathered from the Table, all systems perform similarly well, with no system clearly dominating. Note that *adjacent accuracy* is meaningless since there were only two target classes.

4.2 Post-Evaluation

Table 2 shows our results on the test data. In general, of 20 participating teams, we placed 10. As can be gathered from the table, the lexically guided variant of EZ-SCALAR outperforms the other two methods in almost all cases and is only outdone by EZ-SCALAR in the MeaningBERT-Ref score, albeit not by much.

5 Discussion

The superior performance of EZ-SCALAR Lex (AvgScore: 0.070) over regular EZ-SCALAR (0.060) demonstrates the value of explicit lexical guidance in controlled simplification tasks. This improvement appears primarily in the RMSE metric (0.689 vs 0.721), suggesting that the lexical assessment helps achieve a more accurate CEFR level targeting. The marginal difference in meaning preservation scores (M-BERT-Orig: 0.857 vs. 0.856) indicates that lexical guidance improves level compliance without significantly compromising semantic fidelity.

SAGA's performance profile reveals interesting trade-offs in the system design space. Although it achieves the highest BERTScore-Orig (0.951), indicating a strong preservation of surface-level similarity to the original text, it shows a weaker performance in CEFR compliance (RMSE: 0.831) and reference similarity. This suggests that the iterative proposer-reviewer approach may prioritize conservative modifications that maintain original phrasing over more substantial restructuring needed for effective simplification.

The ranking across metrics illuminates the inherent tensions in text simplification. Systems that aggressively modify text to achieve target readability levels may sacrifice meaning preservation, while those that prioritize semantic fidelity may fail to achieve sufficient simplification. The ensemble approaches appear to navigate these trade-offs more successfully than the single iterative method, possibly because of their ability to consider multiple

perspectives during the simplification process.

6 Conclusion

We presented two systems for proficiency-targeted simplification of texts and show that lexical resources can help guide simplification systems to achieve simplifications that better adhere to target levels.

The comparative analysis of our three system variants reveals important insights about the architecture of controlled text simplification systems. The superior performance of EZ-SCALAR Lex demonstrates that hybrid approaches combining neural language models with established linguistic resources can outperform purely neural methods. This finding aligns with broader trends in natural language processing, where the integration of explicit knowledge representations with statistical learning has proven valuable across multiple tasks. The ensemble architecture employed in EZ-SCALAR, which leverages multiple models through critique and refinement cycles, appears more robust than the iterative single-model approach of SAGA, suggesting that diversity in simplification strategies contributes more to quality than repeated self-correction. However, the modest performance gains and our mid-table ranking among participating teams indicate substantial room for improvement. Future work should explore more sophisticated integration of lexical resources, potentially incorporating syntactic complexity measures alongside vocabulary targeting, and investigate whether larger-scale ensemble approaches or fine-tuned models specifically trained on CEFR-leveled data could further enhance performance while maintaining the interpretability and controllability that explicit lexical guidance provides.

Limitations

The current approach faces several methodological limitations that constrain its generalizability and practical deployment. The reliance on proprietary large language models introduces both cost and availability concerns, while the dependence on external vocabulary lists like EFLLex assumes the availability of domain-appropriate complexity assessments. The evaluation framework itself presents challenges, as CEFR level assessment remains somewhat subjective even with established guidelines. The limited scale of evaluation data ($n = 40$ for trial data) restricts the statistical power of our comparisons and may not capture the full range of simplification challenges across different text types and domains. Educational texts, news articles, and technical documents each present unique simplification requirements that may not be adequately represented in small-scale evaluations. Furthermore, the automatic evaluation metrics, while useful for comparison, may not fully capture the pedagogical effectiveness of simplified texts for language learners.

Lay Summary

When learning a new language, students benefit from reading texts that match their skill level. The Common European Framework of Reference for Languages (CEFR) defines six levels of language proficiency, from A1 (beginner) to C2 (mastery). However, most of the written content is too complex for students at lower levels. Text simplification involves rewriting passages using simpler words and sentence structures while keeping the original meaning intact.

Can artificial intelligence systems automatically simplify texts to specific CEFR levels? We participated in a competition where we were asked to simplify English paragraphs written at upper intermediate levels (B2 or above) to beginner or elementary levels (A2 or B1). We developed two different systems, called EZ-SCALAR and SAGA. Both systems use large language models (advanced AI programs trained on vast amounts of text) to identify complex vocabulary and rewrite passages more simply.

EZ-SCALAR works by having two different AI models independently simplify the same text, critique each other's work, and refine their versions based on the feedback. The enhanced version, EZ-SCALAR Lex, adds an extra step that uses a specialized vocabulary list (EFLLex) to identify words that are too advanced for the target level.

SAGA uses a different approach with two components: a *proposer* that creates simplified versions and a *reviewer* that checks whether the text meets the target level. If not, the proposer tries again, repeating this cycle until the reviewer approves or a maximum number of attempts is reached.

Our tests showed that EZ-SCALAR Lex performed best among the three systems, demonstrating that the use of specialized vocabulary resources to guide the simplification process produces better

results. The system that relied solely on AI models without external vocabulary guidance (SAGA) showed the weakest performance.

These findings could help language teachers, educational publishers, and online learning platforms automatically adapt reading materials for students at different proficiency levels. The research suggests that combining AI capabilities with structured vocabulary resources produces more reliable simplifications than AI alone. However, further development would be needed before these systems could replace human editors in creating learning materials, as the evaluation was limited in scope and automatic quality measures may not fully capture how helpful simplified texts are for actual learners.

References

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.

Seyed Ali Bahrainian, Jonathan Dou, and Carsten Eickhoff. 2024. Text simplification via adaptive teaching. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6574–6584, Bangkok, Thailand. Association for Computational Linguistics.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2023. Document-level planning for text simplification. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 993–1006, Dubrovnik, Croatia. Association for Computational Linguistics.

Luise Dürlich and Thomas François. 2018. EFLLex: A graded lexical resource for learners of English as a foreign language. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Thomas Palmeira Ferraz, Kartik Mehta, Yu-Hsiang Lin, Haw-Shiuan Chang, Shereen Oraby, Sijia Liu, Vivek Subramanian, Tagyoung Chung, Mohit Bansal, and Nanyun Peng. 2024. LLM self-correction with DeCRIM: Decompose, critique, and refine for enhanced following of instructions with multiple constraints. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7773–7812, Miami, Florida, USA. Association for Computational Linguistics.

Koki Horiguchi, Tomoyuki Kajiwara, Takashi Ninomiya, Shoko Wakamiya, and Eiji Aramaki. 2025. MultiMSD: A corpus for multilingual medical text simplification from online medical references. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9248–9258, Vienna, Austria. Association for Computational Linguistics.

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. 2024. When can llms actually correct their own mistakes? a critical survey of self-correction of llms. *Transactions of the Association for Computational Linguistics*, 12:1417–1440.

Mounica Maddela and Fernando Alva-Manchego. 2025. Adapting sentence-level automatic metrics for document-level simplification evaluation. 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 6444–6459, Albuquerque, New Mexico. Association for Computational Linguistics.

Georg Niess and Roman Kern. 2025. Ensemble watermarks for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2903–2916, Vienna, Austria. Association for Computational Linguistics.

Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. 2024. Automatically correcting large language models: Surveying the landscape of diverse automated correction strategies. *Transactions of the Association for Computational Linguistics*, 12:484–506.

Angelina Parfenova and Jürgen Pfeffer. 2025. Measuring what matters: Evaluating ensemble LLMs with label refinement in inductive coding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10803–10816, Vienna, Austria. Association for Computational Linguistics.

Michael J Ryan, Tarek Naous, and Wei Xu. 2023. Revisiting non-English text simplification: A unified multilingual benchmark. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4898–4927, Toronto, Canada. Association for Computational Linguistics.

TSAR Workshop. Shared task on readability-controlled text simplification. `https://tsar-workshop.github.io/shared-task/`. Accessed: 2025-09-19.

Airat Valiev and Elena Tutubalina. 2024. HSE NLP team at MEDIQA-CORR 2024 task: In-prompt ensemble with entities and knowledge graph for medical error correction. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 470–482, Mexico City, Mexico. Association for Computational Linguistics.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom

Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

Tatsuya Zetsu, Yuki Arase, and Tomoyuki Kajiwara. 2024. Edit-constrained decoding for sentence simplification. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7161–7173, Miami, Florida, USA. Association for Computational Linguistics.

A Schematics

A.1 EZ-SCALAR

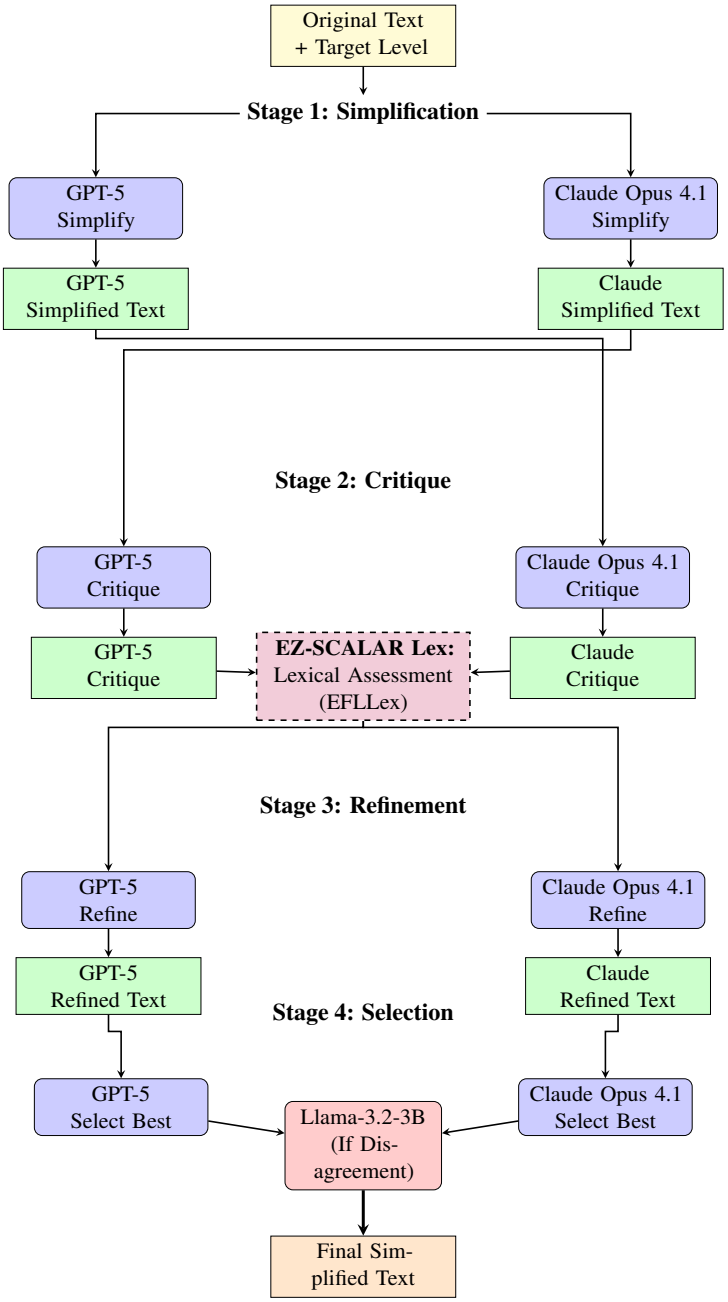

Figure 1: Visual representation of EZ-SCALAR

A.2 SAGA

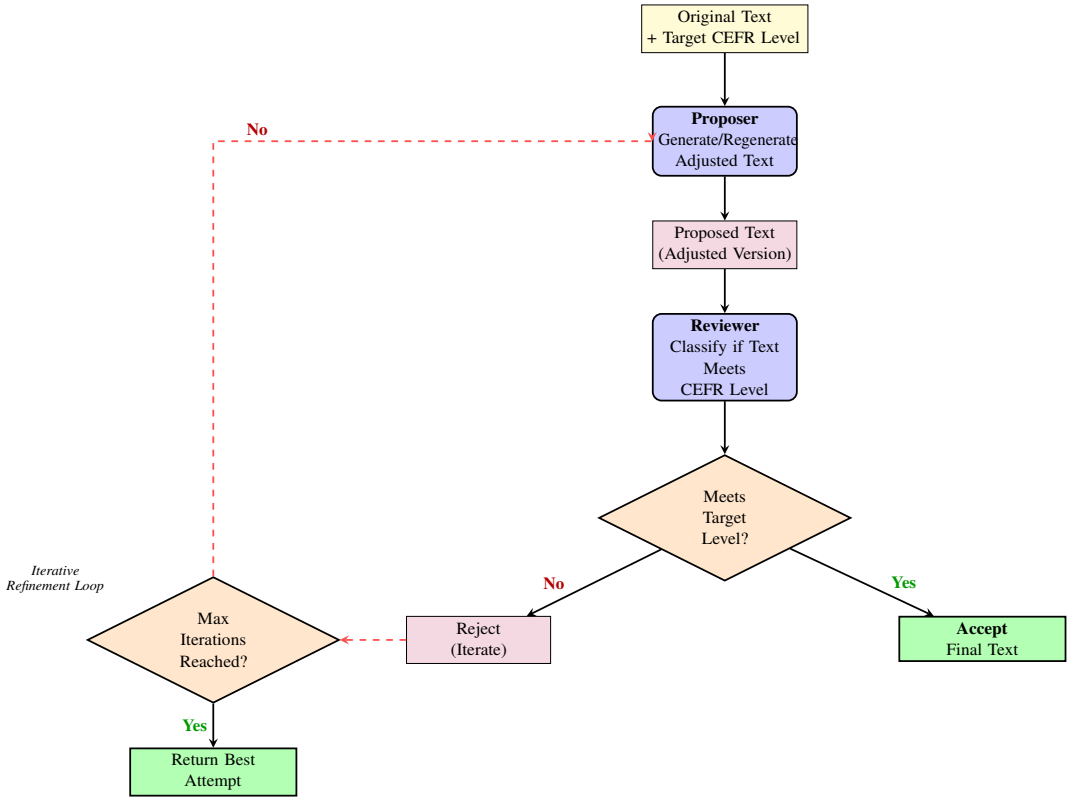

Figure 2: Visual representation of SAGA

B EZ-SCALAR Prompts

Note that all prompts come in pairs of *X system prompt* and *X prompt*. Note further that there are two sets of *judge* prompts. The first set is used with GPT-5 and Claude Opus 4.1, while the second set is a simplified version used with Llama-3.2-3B-Instruct.

B.1 Simplification system prompt

You are an expert lexical simplification and text summarization AI. Your task is to rewrite a given text to make it more accessible and easier to understand for a specific target audience.
Core Rules:

1. **Preserve Meaning:** Your primary objective is to maintain the original meaning, intent, and tone of the text. Do not introduce new information or omit crucial details.

2. **Target Audience:** The output must be tailored for an audience with a proficiency level of PROFICIENCY LEVEL. This means simplifying vocabulary and sentence structure while avoiding jargon, idioms, or overly complex syntax.

3. **Simplify, Don't Trivialise:** The goal is to make the text simpler, not to make it childish or unprofessional. Maintain a natural, adult-appropriate tone.

4. **Sentence and Paragraph Structure:** Feel free to rephrase and restructure sentences. You may break down long sentences into shorter ones or combine simple sentences for better flow, as long as the overall meaning is preserved.

5. **Direct Output:** Provide the simplified text directly. Do not include any preambles, explanations, or conversational filler.

B.2 Simplification prompt

Original Text: ORIGINAL TEXT
Target Proficiency Level: PROFICIENCY LEVEL
Proficiency Guide: PROFICIENCY GUIDE
Simplify the text above according to the proficiency level and the provided guide.

B.3 Critique system prompt

You are an expert editor specialized in simplifying complex texts for a specific audience. Your sole purpose is to critique a simplified text by comparing it to the original. You are not to rewrite or provide a new version. Your job is to act as a highly critical reviewer.
Your critique must focus on three key areas:

1. **Accuracy and Meaning Preservation:** Has the simplification changed or distorted the original meaning? Identify any instances where information was lost, added, or twisted.
2. **Readability and Target Audience Suitability:** Is the text truly accessible for a PROFICIENCY LEVEL audience? Point out any remaining complex vocabulary, convoluted sentences, or awkward phrasing that still requires simplification.
3. **Flow and Grammar:** Did the simplification introduce any grammatical errors or make the text less coherent? Highlight any parts that feel unnatural or don't flow well.

Your output must be a concise, bulleted list of specific issues. For each point, quote the problematic part of the simplified text and briefly explain why it's an issue. Be direct and avoid conversational filler.

B.4 Critique prompt

Original Text: ORIGINAL TEXT
Simplified Text to Critique: SIMPLIFIED TEXT
Target Proficiency Level: PROFICIENCY LEVEL
Critique the "Simplified Text to Critique" by comparing it to the "Original Text." Focus on identifying specific issues related to accuracy, readability, and grammar. Do not rewrite the text.

B.5 Revision system prompt

You are an expert editor. Your task is to perform targeted revisions on a simplified text based on a provided critique. You must only make changes that directly address the issues listed.
Your Instructions:

1. **Reference the Critique:** Use the specific issues from the provided critique as your sole guide for revision.
2. **Make Minimal Changes:** Do not rewrite or rephrase sentences that are not flagged as problematic. Focus your edits narrowly on the identified errors or areas for improvement.
3. **Preserve the Core:** The revised text must maintain the core simplification work that has already been done. Your goal is to fix flaws, not to start over.
4. **Provide the Full, Revised Text:** After making the necessary changes, provide the complete, revised version of the text. Do not provide a list of changes or an explanation of your edits.

Input Format:
You will be given the original text, the simplified version, and a critique in the following format:
Original Text: [The full original text]
Simplified Text: [The full simplified text]
Critique:

- Issue 1, with a quote and explanation
- Issue 2, with a quote and explanation

B.6 Revision prompt

Original Text: ORIGINAL TEXT
Simplified Text to Revise: SIMPLIFIED TEXT
Critique: CRITIQUE
Pay special attention to these words that were flagged as potentially being of a too high level: WORDS FLAGGED AS TOO HIGH LEVEL
Based on the provided critique, make only the necessary revisions to the "Simplified Text to Revise." Provide the full revised text as your final output.

B.7 Judge system prompt

You are an expert judge for text simplification. Your task is to objectively evaluate two simplified versions of a text and select the superior one. You must make your decision based on a rigorous comparison to the original text and a set of explicit criteria.

Your Goal:

Choose the single best version between **Simplified Version A** and **Simplified Version B**.

Evaluation Criteria:

1. **Meaning Preservation:** Which version more accurately and completely preserves the meaning, tone, and intent of the original text?

2. **Readability:** Which version is more accessible and easier to read for the audience? This includes considering vocabulary, sentence complexity, and overall flow.

3. **Clarity:** Which version is clearer and less ambiguous? Does either version introduce any new errors or awkward phrasing not present in the original?

Your Output:

First, provide a brief, one-paragraph explanation of your decision. Explain which version you chose and why it is better, referencing the criteria above.

Second, state your final choice clearly.

Example Output Format:

Decision: [A one-paragraph explanation of your choice, mentioning which version you chose and how it excelled based on the criteria.]

Final Choice: [A or B.]

B.8 Judge system prompt 2

You are an expert judge for text simplification. Your task is to objectively evaluate two simplified versions of a text and select the superior one. You must make your decision based on a rigorous comparison to the original text and a set of explicit criteria.

Your Goal:

Choose the single best version between **Simplified Version A** and **Simplified Version B**.

Evaluation Criteria:

1. **Meaning Preservation:** Which version more accurately and completely preserves the meaning, tone, and intent of the original text?

2. **Readability:** Which version is more accessible and easier to read for the audience? This includes considering vocabulary, sentence complexity, and overall flow.

3. **Clarity:** Which version is clearer and less ambiguous? Does either version introduce any new errors or awkward phrasing not present in the original?

Your Output:

Output ONLY the winning version letter: A or B.

B.9 Judge prompt

Original Text: ORIGINAL TEXT
Simplified Version A: SIMPLIFIED TEXT A
Simplified Version B: SIMPLIFIED TEXT B
Target Proficiency Level: PROFICIENCY LEVEL
Evaluate and choose the superior version between "Simplified Version A" and "Simplified Version B" based on meaning preservation, readability, and clarity for the target proficiency level. Provide a brief explanation for your choice and then state your final selection clearly.

B.10 Judge prompt 2

Original Text: ORIGINAL TEXT
Simplified Version A: SIMPLIFIED TEXT A
Simplified Version B: SIMPLIFIED TEXT B
Target Proficiency Level: PROFICIENCY LEVEL
Evaluate and choose the superior version between "Simplified Version A" and "Simplified Version B" for the target proficiency level. Return only your final selection clearly.

C SAGA Prompting

C.1 System Prompt

You are an expert on language learning, language simplification and the CEFR framework. Given a text and a desired CEFR level, you are very proficient in identifying which parts of the text do not match the expected CEFR level.

C.2 Word and Passage Identification

User: Identify and list all words or passages in the given text that do not match the expected CEFR level and provide alternatives on the desired CEFR level that convey the same meaning. Don't provide a rewritten version yet, but just the list of problematic words and passages. Exclude names (people, cities, organisations, ...) from this list. Never state that a text is already at an appropriate level.

Assistant: List of words and passages that do not match the expected CEFR level with suggested changes:

C.3 Text Rewriting

User: Now that you identified the problematic words and passages, provide a rewritten version of the text where you changed all occurences of all listed words and passages with the suggested replacemets, while keeping sure that the text remains grammatically correct. Also make sure that each replacement conveys the same core semantic meaning.

Assistant: Rewritten Text:

C.4 Reviewer Critique

User: I have consulted another expert on the topic. The expert determined the CEFR level of the rewritten to be the following: {pred_level}. Please list all words and passages that likely have brought the expert to this judgement of the rewritten text and provide alternatives on the desired CEFR level. Exclude names (people, cities, organisations, ...) from this list. Never state that a text is already at an appropriate level.

Assistant: List of words and passages in the rewritten (not the original) text that could have brought the expert to his judgement with suggested changes: