UoL-UPF at TSAR 2025 Shared Task: A Generate-and-Select Approach for Readability-Controlled Text Simplification

Akio Hayakawa[1]   **Nouran Khallaf**[2]   **Serge Sharoff**[2]   **Horacio Saggion**[1]

[1]Universitat Pompeu Fabra   [2]University of Leeds
{akio.hayakawa, horacio.saggion}@upf.edu
{N.Khallaf, S.Sharoff}@leeds.ac.uk

Abstract

The TSAR 2025 Shared Task on Readability-Controlled Text Simplification focuses on simplifying English paragraphs written at an advanced level (B2 or higher) and rewriting them to target CEFR levels (A2 or B1). The challenge is to reduce linguistic complexity without sacrificing coherence or meaning. We developed three complementary approaches based on large language models (LLMs). The first approach (Run 1) generates a diverse set of paragraph-level simplifications. It then applies filters to enforce CEFR alignment, preserve meaning, and encourage diversity, and finally selects the candidates with the lowest perceived risk. The second (Run 2) performs simplification at the sentence level, combining structured prompting, coreference resolution, and explainable AI techniques to highlight influential phrases, with candidate selection guided by automatic and LLM-based judges. The third hybrid approach (Run 3) integrates both strategies by pooling paragraph- and sentence-level simplifications, and subsequently applying the identical filtering and selection architecture used in Run 1. In the official TSAR evaluation, the hybrid system ranked **2nd overall**, while its component systems also achieved competitive results.

1 Introduction

Text Simplification aims to make complex texts more accessible to a broad audience, including language learners and individuals with reading difficulties (Saggion, 2017; Al-Thanyyan and Azmi, 2021). However, many traditional approaches fail to meet the diverse needs of readers at different proficiency levels. To address this, the field has moved towards targeted simplification, which aims to adapt the complexity of a text to a specific reader's needs, rather than just simplifying it for a general audience (Barayan et al., 2025; Säuberli et al., 2024). This requires defining specific proficiency targets, and

the Common European Framework of Reference for Languages (CEFR) has been widely used for this purpose (Imperial et al., 2025). Also, the majority of text simplification research has focused on sentence-level, while largely overlooking the more practical scenario of paragraph-level simplification. The TSAR 2025 Shared Task on Readability-Controlled Text Simplification is situated within this context, challenging participants to simplify paragraphs originally at B2 level or above to target levels of A2 and B1 (Alva-Manchego et al., 2025).

In this paper, we propose and validate a Generate-and-Select approach that does not rely on a single best prompt, model, or simplification strategy. Our primary goal was to achieve a high score on a key evaluation metric: **similarity to the reference text**. The official evaluation, conducted only automatically, was based on three metrics: CEFR compliance, output-to-original similarity (Meaning Preservation), and output-to-reference similarity. While the first two could be calculated by participants themselves, the reference texts were not provided. Our system therefore aimed for a high output-to-reference similarity.

To achieve this, we developed a powerful generate-and-select pipeline based on paragraph-level simplification (Run 1) as our core approach. This system first generates a diverse set of candidates and then filtered to create a high-quality candidate pool for Minimum Bayes Risk (MBR) decoding (Bickel et al., 2005) to select the optimal output. As demonstrated by Heineman et al. (2024), the diversity of candidates is crucial for enhancing the quality of MBR decoding. To further improve its performance, we introduced a sentence-level system (Run 2). While weaker on its own, this secondary system successfully injected structural diversity into our candidate pool. Our final, hybrid system (Run 3) combines the candidate pool from both Run 1 and Run 2. It then processes this combined pool using the same pipeline as Run 1 to

select the optimal output.

Our approach proved highly effective in the shared task. Among 48 submissions from 20 international teams, our hybrid system (Run 3) and core system (Run 1) placed **2nd and 3rd overall**. Notably, **Run 3 and 1 ranked 1st and 2nd on the reference text similarity** respectively, confirming the success of our primary objective.

However, our success also revealed an inherent limitation of the evaluation metric we focused on optimizing. Our case study highlights that while the metric is designed to capture deep semantic similarity, its scores can still be influenced by surface-level features. This can be misleading, as lexical overlap can sometimes outweigh semantic factuality in the score.

The main contributions of this paper are:

- We present a Generate-and-Select pipeline that successfully maximizes reference similarity.
- We demonstrate that even a weak system can contribute the diversity needed for a powerful selection pipeline.
- We analyse the limitations of the evaluation metric we focused on optimizing.

The experimental setup is available on GitHub.[1]

2 Our pipeline

Our submission consists of three systems (Runs 1-3). Our core approach, which achieved 3rd place overall, is presented as Run 1. While our primary objective is to achieve a high output-to-reference similarity, we also aim to attain satisfactory scores in other metrics, namely CEFR compliance and meaning preservation.

2.1 Run 1: Paragraph-Level MBR System

Run 1 is our primary system, designed to maximize the similarity between system outputs and reference texts, through a multi-stage pipeline. As shown in Figure 1, the core approach is a three-stage process. We first generate a diverse set of candidates, and then select a high-quality subset by applying CEFR and Meaning Preservation filtering. Finally, we apply MBR decoding to select the output with the lowest risk.

2.1.1 Diverse Candidate Generation

The process starts with generating a large set of initial simplification candidates for each source paragraph and its corresponding target CEFR level. To
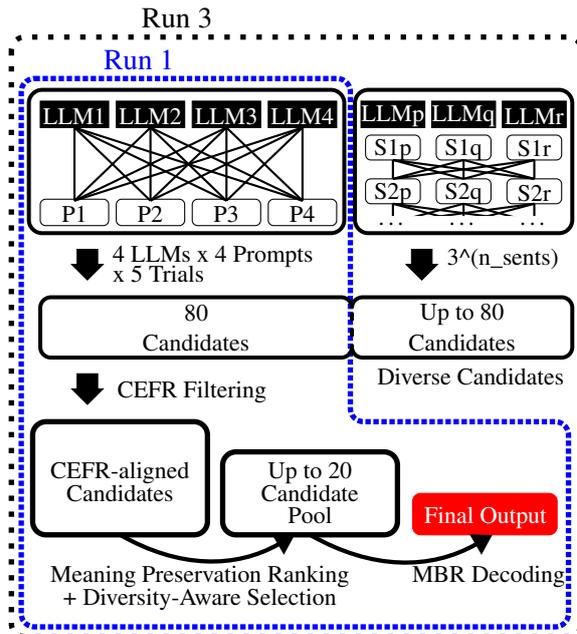

Figure 1: System Architecture of Run 1 and 3.

ensure a rich and varied candidate pool, this generation process employs two key diversity strategies: multi-prompting and multi-model.

- **Multi-Prompting**: We prepare four types of prompts, with three of them automatically generated by an LLM. Our prompts include two inductive prompts derived from trial data, a deductive prompt based on CEFR-adapted simplification rules, and a standard few-shot prompt. (See Appendix A for the details.)
- **Multi-Model**: The prompts above are run across four auto-regressive large language models (LLMs), GPT-4.1-mini,[2] gpt-oss-20b (OpenAI, 2025), Gemma-3-4b-it (Gemma, 2025), and Qwen-2.5-14b-it (Qwen, 2025), to capture the unique simplification tendencies of each model.

For each combination of prompt and LLM, we performed five simplification trials, using five separate API calls or five different seeds. As a result, we generated 80 candidates per simplification instance (4 LLMs x 4 prompts x 5 trials). See Appendix C for the hyperparameter settings.

2.1.2 Candidate Pool Construction

After the generation stage, we filter, rank, and select from the initial set of candidates. This process creates an optimized candidate pool of up to 20 simplifications for MBR decoding.

---

[1] https://github.com/ahaya3776/tsar2025sharedtask-uol-upf

[2] https://openai.com/index/gpt-4-1/

1. **CEFR Filtering**: First, we label the CEFR level (A1, A2, B1, B2, C1, and C2) for all candidates and obtain the minimum difference from the target CEFR level. Given the large number of candidates, this minimum difference is almost always zero (i.e., at least one candidate matches the target CEFR level). We then retain only the candidates that have this minimum difference. CEFR levels are labeled using classification models used in the official shared task evaluation.

2. **Meaning Preservation Ranking**: The remaining CEFR-compliant candidates are ranked in their semantic similarity to the original source paragraph. We use MeaningBERT (Beauchemin et al., 2023) following the official evaluation.

3. **Diversity-Aware Selection**: From this ranked list, we build the final pool with a maximum size of 20. We select candidates primarily based on the previous ranking. However, to maximize the benefits of MBR decoding, which requires a diverse candidate pool (Heineman et al., 2024), we apply a filter to ensure structural diversity. A candidate is added to the pool only if its BLEU (Papineni et al., 2002) against every candidate already in the pool is below a threshold of 0.5.

2.1.3 MBR Decoding

Finally, we apply MBR decoding to the constructed pool. MBR selects the single candidate that maximizes the expected utility function against all other candidates in the set. For the utility function, we again use MeaningBERT, measuring the pairwise similarity between candidates. The candidate with the highest average similarity score against its other candidates is selected as the final output. The final output $\hat{y}_{MBR}$ can be expressed as:

$$\hat{y}_{MBR} = \underset{y \in \mathcal{H}}{\operatorname{argmax}}(\mathbb{E}_{\mathcal{H}}[\mathbb{E}_{y' \in \mathcal{H}}[u(y, y')]]), \quad (1)$$

where $\mathcal{H}$ is a candidate pool and $u(y, y')$ is a utility function, defined as $\operatorname{MeaningBERT}(y, y')$.

2.2 Run 2: Sentence-level Simplification

Our second system approaches the task at the sentence level. Prior work has shown that long, coreferential sentences with dense terminology are a key source of difficulty for readers and are best addressed through targeted edits rather than global rewrites (Siddharthan, 2006; Shardlow, 2014; Štajner and Popović, 2016; Barayan et al., 2025). Run 2 therefore investigates whether *explicit linguistic*

control that applied locally at the sentence level, can better align outputs with CEFR levels while preserving meaning (for system architecture see Appendix E). By simplifying sentences independently, while still highlighting the most important phrases, we aim to produce outputs that are both controlled and interpretable. Run 2 consists of the following steps.

Preprocessing. Each paragraph is first segmented into sentences and normalised for coreference. We replace ambiguous pronominal references (e.g., *he, she, they, it*) with their antecedents using AllenNLP's coreference system (Lee et al., 2017) and the spaCy-compatible coref module (Honnibal et al., 2020). This produces a list of self-contained sentences that can be simplified *independently*.

Highlighting influential phrases. To identify which parts of a sentence contribute most to linguistic complexity, we apply **Integrated Gradients (IG)** (Sundarararajan et al., 2017). We apply Captum's `LayerIntegratedGradients` (Miglani et al., 2023) over the embedding layer of a sentence-based CEFR classifier (Barayan et al., 2025), using a padded baseline sequence and integrating gradients with respect to the "complexity" logit. Token-level attribution scores are aggregated into multi-word phrases (NP, VP, ADJP, PP) using spaCy chunks. The top-K phrases (default $K=6$) are retained by absolute score. These influential phrases are exported as $(\text{type}, \text{phrase}, \text{score})$ triples and injected into the simplification prompt (see Appendix B.1). This allows the LLM to focus on which terms to simplify or gloss.

The same influential phrases have another role in the evaluator step, in which the metric verifies whether these spans are preserved in the simplified output. In this way, IG attributions serve a dual purpose: guiding generation and informing evaluation.

Simplification strategies. We guide the models with strategies inspired by intralingual translation and Easy-to-Read (E2R) English (Khallaf et al., 2025). These include **explanation** (adding glosses), **modulation** (one idea per sentence), **synonymy** (simpler words), **syntactic changes** (splitting clauses), and **omission** (dropping non-essential details).

Prompting and candidate generation. We prompt three LLMs, LLaMA-3-8B (Dubey et al.,

2024), GPT-4o (OpenAI, 2023), and Mistral-7B (Jiang et al., 2023), to generate simplifications for CEFR levels A1, A2, and B1 in a single response. Prompts enforce constraints on meaning preservation, correctness of entities and numbers, readability (shorter sentences, simpler words), and strict formatting with explicit level tags (see Appendix B.1).

Automatic and hybrid judging. Candidate outputs are scored by an automatic judge that integrates eight complementary signals (see Table 7 in Appendix F). These include semantic similarity based on sentence embeddings and entailment (Williams et al., 2018), key-phrase coverage from IG attributions, entity and number fidelity using spaCy (Honnibal et al., 2020), readability targets derived from average sentence length (ASL) and Flesch Reading Ease (Flesch, 1948), lexical simplification (syllable reduction), fluency via language model perplexity (Jurafsky and Martin, 2023), compression ratio, and sentence/format control.

We combine heterogeneous metrics with a *weighted geometric mean*, which is widely used in multi-criteria evaluation (Krejčí and Stoklasa, 2018; Herbold, 2020). When two candidates score within a small margin, we invoke a **Hybrid Auto+LLM (HAI) judge**, which queries a second LLM (GPT-4o or LLaMA) to make a pairwise choice with justification. We pass the original, target level, and top-K candidates (prefiltered by the auto judge) to a second LLM (GPT-4o or LLaMA) that returns a winner index and reason (see Appendix B.2). After simplification, sentences are re-stitched into the level-tagged block (<B1>, <A2>, <A1>)

2.3 Run 3: Hybrid MBR System

Our best-performing system, Run 3, uses the same pipeline as Run 1 but starts with a more diverse set of initial candidates from Run 2. In addition to 80 candidates generated in Run 1, we incorporate candidates based on sentence-level simplification in Run 2. As shown in Figure 1, we generate candidates based on Run 2 by concatenating sentence-level simplifications. For each sentence in an original paragraph, three simplified sentences are generated by three different LLMs. The combination of simplified sentences result in $3^{n\_sentences}$ potential paragraph variants, from which we randomly sample up to 80 candidates. Among this combined set of up to 160 candidates, the final output is selected through the identical process described for Run 1.

Team	CEFR RMSE	Sim Orig	Sim Ref	Total Rank
EhiMeNLP	0.000	.902	.845	1
UoL-UPF (3)	0.000	.856	.857	2
UoL-UPF (1)	0.000	.849	.856	3
HIT-YOU	0.158	.852	.835	4
Archaeology	0.122	.779	.804	11
ounlp	0.755	.855	.849	14
SQUREL	1.153	.979	.819	23
UoL-UPF (2)	0.693	.808	.827	-

Table 1: Representative results from 44 runs from 20 teams. The best performance for each metric is shown in red. Run 2 is an unofficial result due to parsing error, and its estimated rank is around 20th.

	A2		B1	
Model	Num.	Sim	Num.	Sim
GPT-4.1-mini	24	.841	13	.865
gpt-oss-20b	31	.831	17	.902
Gemma-3-4b	16	.840	12	.862
Qwen-2.5-14b	26	.862	36	.877
Sentence-lv	3	.730	22	.860

	A2		B1	
Prompt	Num.	Sim	Num.	Sim
Prompt 1	19	.839	20	.872
Prompt 2	30	.838	15	.908
Prompt 3	24	.831	23	.867
Prompt 4	24	.866	20	.874
Sentence-lv	3	.730	22	.860

Table 2: Distribution of models and prompts selected as a final candidate in Run 3 with output-to-reference similarity scores by MeaningBERT.

3 Results and Discussions

Table 1 shows the official results of the shared task. The hybrid system (Run 3) is ranked 2nd, while the core system (Run 1) is 3rd overall. Furthermore, our systems placed 1st (tied, full marks) on CEFR alignment, and 1st and 2nd on output-to-reference similarity. This result confirms the success of our pipeline combining filtering and MBR decoding, thereby achieving the high output-to-reference similarity while maintaining other metrics.

Table 2 demonstrates the distribution of selected candidates for Run 3, categorized by their source. The selections were generally distributed evenly across target levels and our various prompts, models, and granularities. The only exception is

Ablation	A2		B1	
	Orig	Ref	Orig	Ref
Run 3	.836	.840	.876	.874
w/o Sent. lv (≡ Run 1)	.824	.837	.874	.875
w/o MPR, DAS, MBR	.756	.779	.817	.822
w/o MPR, DAS	.815	.830	.850	.858
w/o DAS	.849	.834	.891	.869
w/o MBR (Random)	.789	.793	.841	.832
w/o MBR (Highest MP)	.896	.814	.919	.858
w/ smaller MBR (size=10)	.853	.838	.888	.873

Table 3: MeaningBERT scores between outputs and original (Orig) and reference (Ref), as an ablation study for processes after the CEFR filtering. MPR and DAS refers to Meaning Preservation Ranking and Diversity-Aware Selection, respectively.

sentence-level approach for the A2 target. This implies that adding explanations, often observed in the simplification to lower proficiency levels, is hard to achieve via sentence-level approach. This overall diversity was the key to the success of our MBR-based selection pipeline.

Furthermore, we conducted ablation study shown in Table 3. As we described, final outputs are selected through Meaning Preservation Ranking, Diversity-Aware Selection, and MBR decoding after the CEFR filtering. The study shows that each of these steps contributed to improve output-to-reference similarity. Notably, MBR decoding boosted it up, while increasing the candidate pool size produced only a negligible gain.

This success also highlights an important characteristic of our method. Figure 2 illustrates the MeaningBERT scores distribution of CEFR-aligning candidates for one example instance. While the final output shows the highest output-to-original similarity, several candidates show higher output-to-reference similarity. This observation confirms that MBR decoding is designed to minimize the risk of selecting a low-scoring candidate, not to select one with the maximum expected score. As a result, final outputs are often conservative.

Despite prioritizing output-to-reference similarity, we acknowledge that over-reliance on this metric can be problematic. Our qualitative analysis shows limited agreement between scores and human judgments. Specifically, instances containing semantic errors or complex vocabulary (yellow in the scatter plot) are often over-evaluated by the metric when they are structurally similar to the reference. On the other hand, structure changes, such as sentence splitting, are penalized even if beneficial.
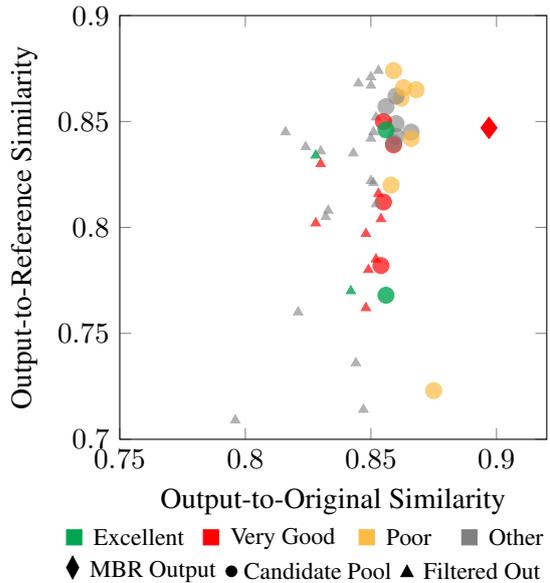

Figure 2: Scatter plot for CEFR-aligned candidates of a single instance. Axes represent similarity scores between output and original/reference. Circles are ones selected as candidate pool, and the diamond is the final output through MBR decoding. Colors align with Table 5, Table 6 in which we manually judged simplification quality.

Our case study supports that MeaningBERT often fail to capture the value of features such as sentence splitting, synonym choice, and moral or pragmatic clarity, rewarding surface overlap instead of genuine accessibility (Barayan et al., 2025). We provide full analysis in Appendix D.

4 Conclusion

In this paper, we presented our Generate-and-Select framework for the TSAR 2025 shared task, which achieved 2nd and 3rd place overall. Our core approach utilized a diverse candidate pool from multiple LLMs and prompts, with MBR decoding for robust selection.

Our primary contribution is demonstrating that our Generate-and-Select framework is highly effective. We showed that its strength lies in prioritizing the diversity of candidates, which allowed even a weaker system (our sentence-level Run 2) to make a contribution to the final performance by injecting variety.

Finally, our analysis shows that while our pipeline is robust, its limitation in a single-reference context highlights the need for selection methods that can better handle unpredictable simplifications.

Lay Summary

UoL-UPF team participated in the TSAR 2025 Shared Task. The goal of this shared task was to rewrite difficult English texts into simple texts at a specific level.

We tried an idea we call Generate-and-Select approach. In this approach, first, we used LLMs to generate many versions of simple texts. We used different LLMs and prompts, so there were a lot of options to choose from. This variety was a key part of our idea. Next, we selected the best option from these simple texts. We built a system to check all the simple texts. This system had some filtering processes. For example, one filter only selected texts that were similar to original difficult texts. After these filtering processes, we only had high-quality options. Finally, from these high-quality options, we selected the lowest-risk option as a final result.

Our system performed very well, and was ranked 2nd place out of 48 systems. This great result showed that our idea was a good one. Through this project, we learned some very important things. It is true that our generate-and-select approach works well, especially when the quality of generated texts is judged by computer. However, we cannot always trust computer judge. In our study, some simple texts were good by computer judge, but not by human judge.

Limitations

The primary limitation of this work is its reliance on diverse set of generation. While the LLMs we employed are relatively small-scaled and thus do not require excessive computational resources, the time and cost associated with obtaining the final outputs cannot be disregarded. Therefore, our generate-and-select framework would be unsuitable for real-time text simplification.

Also, this shared task relies on automatic evaluation metrics. While our system achieved high scores, we did not conduct a manual evaluation with human participants to confirm whether the outputs are genuinely more readable and understandable for the target readers. Such manual evaluation, with Likert scoring or reading comprehension questions, would be necessary to validate the real-world effectiveness of our simplifications.

References

Suha S Al-Thanyyan and Aqil M Azmi. 2021. Automated text simplification: a survey. *ACM Computing Surveys (CSUR)*, 54(2):1–36.

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 (COLING)*, pages 6762–6781. Association for Computational Linguistics.

David Beauchemin, Horacio Saggion, and Richard Khoury. 2023. Meaningbert: assessing meaning preservation between sentences. *Frontiers in Artificial Intelligence*, 6:1223924.

Steffen Bickel, Peter Haider, and Tobias Scheffer. 2005. Predicting sentences using n-gram language models.

In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 193–200, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Abhimanyu Dubey, Rohan Taori, Alexei Baevski, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Rudolf Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

Gemma. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

David Heineman, Yao Dou, and Wei Xu. 2024. Improving minimum Bayes risk decoding with multi-prompt. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22525–22545, Miami, Florida, USA. Association for Computational Linguistics.

Steffen Herbold. 2020. Autorank: A python package for automated ranking of classifiers. *Journal of Open Source Software*, 5(48):2173.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. Software documentation.

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*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *Preprint*, arXiv:2310.06825.

Dan Jurafsky and James H. Martin. 2023. Speech and language processing (3rd ed. draft): Chapter on language modeling. Online draft.

Nouran Khallaf, Carlo Eugeni, and Serge Sharoff. 2025. Reading between the lines: A dataset and a study on why some texts are tougher than others. *Preprint*, arXiv:2501.01796. Published at Writing Aids at the Crossroads of AI, Cognitive Science and NLP (WR-AI-CogS), COLING 2025, Abu Dhabi.

Jana Krejčí and Jan Stoklasa. 2018. Aggregation in the analytic hierarchy process: Why weighted geometric mean should be used instead of weighted arithmetic mean. *Expert Systems with Applications*, 114:97–106.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Vivek Miglani, Aobo Yang, Aram Markosyan, Diego Garcia-Olano, and Narine Kokhlikyan. 2023. Using Captum to explain generative language models. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 165–173, Singapore. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *Preprint*, arXiv:2303.08774.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Qwen. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Horacio Saggion. 2017. *Automatic Text Simplification*, volume 14123 of *Lecture Notes in Computer Science*. Springer.

Andreas Säuberli, Franz Holzknecht, Patrick Haller, Silvana Deilen, Laura Schiffl, Silvia Hansen-Schirra, and Sarah Ebling. 2024. Digital comprehensibility assessment of simplified texts among persons with intellectual disabilities. CHI '24, pages 1–11, New York, NY, USA. Association for Computing Machinery.

Matthew Shardlow. 2014. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*.

Advaith Siddharthan. 2006. Syntactic simplification and text cohesion. *Research on Language and Computation*, 4(1):77–109.

Sanja Štajner and Maja Popović. 2016. Can text simplification improve machine translation? In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*, pages 172–178. European Language Resources Association.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3319–3328, Sydney, Australia. PMLR.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

A Run 1: Prompts for Paragraph-level Simplification

We used four simplification prompts for LLMs. Two of these were based on inductive approach, which involved extracting simplification features from trial data to create instructions as a prompt. To do this, the following prompt was given to GPT-4.1-mini.

> You will be given several pairs of paragraphs. Each pair is composed of an original paragraph and a simplified version for CEFR {lv} readers. Your task is to analyze these pairs to find the general patterns of simplification and write an instruction for LLMs to simplify paragraphs similarly. Include observations on information or phrasing that remains unchanged. Do not include examples that contain text parts in given paragraphs. Only output your final prompt.
>
> Original: {Original Paragraph 1 of the target CEFR level}
> Reference: {Reference Paragraph 1 of the target CEFR level}
> Original: {Original Paragraph 2 of the target CEFR level}
> Reference: {Reference Paragraph 2 of the target CEFR level}
> ...

After several trials, we picked up following two types of prompts for each level with some minor arrangements.

Prompt 1 : A2

> Simplify paragraphs for CEFR A2 readers by following these guidelines:
> 1. Use short, clear sentences with simple grammar structures (mostly present and past simple).
> 2. Replace complex or abstract vocabulary with common, concrete words; explain any necessary technical terms briefly and clearly.
> 3. Remove or reduce detailed numerical data, statistics, or nuanced concepts unless essential; when included, present numbers simply and round if appropriate.
> 4. Avoid idiomatic expressions, figurative language, and complex sentence forms like passive voice or embedded clauses.
> 5. Focus on main ideas and essential facts; omit detailed background information, speculation, or subtle distinctions unless they support comprehension.
> 6. Use explicit cause–effect and temporal connectors (e.g., because, so, but, then, now) to clarify relationships.
> 7. Maintain logical and coherent flow with clear topic introductions and simple sequencing.
> 8. Preserve proper names, key terms, and notable facts that are central to understanding.

> 9. When appropriate, add brief, straightforward definitions or explanations of less familiar concepts.
> 10. Use active voice predominantly and ensure the subject of sentences is clear.
> 11. Replace pronouns that may confuse with explicit nouns where needed.
> 12. Retain the overall meaning and important details but adapt phrasing to be direct and concrete.
> 13. Introduce examples to illustrate points simply, using familiar or relatable contexts.
> 14. Do not assume prior knowledge; present background information in simple terms if required.
> 15. Where opinion or interpretation appears, present it clearly and simply, often using direct statements like " people say" or "some think."
> 16. Use simple punctuation and avoid complex structures such as long lists or parenthetical asides.
> By following these patterns, produce an accessible, easy–to–read version of a paragraph that preserves the core message and key details for A2–level readers. Provide only the simplified paragraph without any explanation or justification.
>
> # Original:
> {Original Paragraph}
> # Simplified:

Prompt 1 : B1

> Simplify paragraphs for CEFR B1 readers by following these guidelines:
> 1. Use simpler vocabulary and expressions: Replace complex or formal words and phrases with more common, everyday alternatives, while keeping the meaning intact.
> 2. Shorten and clarify sentences: Break long, complex sentences into shorter, clearer ones. Use straightforward sentence structures, avoiding passive voice or complicated clauses.
> 3. Explain or define less familiar terms: When necessary, introduce brief explanations or definitions of technical, cultural, or less common concepts within the text to aid understanding.
> 4. Retain key information and facts: Keep all essential data, figures, names, and core ideas from the original text, ensuring the main message is preserved.
> 5. Rephrase for explicitness and clarity: Make implied meanings more explicit, and clarify references to pronouns or abstract concepts.
> 6. Maintain original factual content and sequence: Do not omit major details or reorder information in ways that change the logical flow or significance.
> 7. Use familiar synonyms and phrases: Prefer words and expressions that are frequently used at intermediate English level rather than academic or highly technical language.
> 8. Simplify complex concepts without oversimplifying: Present difficult ideas in more accessible language but avoid losing the nuance or accuracy of the original content.
> 9. Use concrete examples or context where helpful: When abstract concepts might confuse, add brief relatable examples or contextual cues to aid comprehension.
> 10. Preserve unchanged proper nouns and names: Keep names of people, places, events, titles, and specific terms as in the original to maintain accuracy and recognition.
> 11. Avoid idiomatic or culture–specific expressions unless explained: Replace or explain idioms and culturally specific references that might not be understood by B1 learners.
> 12. Retain the original tone and intent as much as possible :** The simplification should respect the author's purpose, tone, and the overall style, aiming for clarity rather than casualness.

In summary, simplify language and sentence structure, clarify meaning, explain or define unfamiliar terms, keep all important facts and details, and ensure the text remains coherent and faithful to the original. Provide only the simplified paragraph without any explanation or justification.

Original:
{Original Paragraph}
Simplified:

Prompt 2 : A2

Simplify paragraphs for CEFR A2 readers by following these guidelines:
1. **Vocabulary and Grammar:**
 – Use very common, everyday words and simple sentence structures.
 – Avoid idioms, metaphors, or abstract expressions.
 – Prefer present tense or simple past tense; avoid complex verb forms.
 – Use short sentences, often one idea per sentence.
2. **Sentence Structure:**
 – Break long, complex sentences into multiple shorter sentences.
 – Use basic conjunctions (and, but, so, because) to connect ideas simply.
 – Avoid passive voice where possible; use active voice instead.
3. **Information Selection and Clarity:**
 – Retain all key factual information from the original paragraph.
 – Remove or rephrase any statistics or figures only if they might confuse the reader, but generally keep numbers with simple explanations.
 – Explain or define any technical terms or names using simple language or familiar examples.
 – Avoid unnecessary detail or background information unless it helps understanding.
4. **Rephrasing and Simplification:**
 – Replace complex nouns or phrases with simpler equivalents or brief explanations.
 – Make implicit information explicit if needed.
 – Use examples or explanations to clarify concepts that might be unfamiliar.
 – Use repetition and restatement to reinforce understanding without changing meaning.
5. **Tone and Style:**
 – Use a neutral, clear, and straightforward tone.
 – Address the reader more directly and simply when appropriate.
 – Keep the original meaning, emphasis, and main points intact.
6. **Preserving Key Proper Nouns and Data:**
 – Keep proper names (people, places, organizations, titles) unchanged but briefly explain their significance if needed.
 – Maintain important dates, measurements, and specific figures, simplifying explanations around them.
7. **Avoid Removing Content Entirely:**
 – Instead of deleting difficult or nuanced content, re–express it in accessible language.
 – Questions or rhetorical devices in the original can be kept but simplified and clarified.
By applying these principles, transform original paragraphs into clear, accessible text suitable for A2–level readers while preserving essential information and intent. Provide only the simplified paragraph without any explanation or justification.

Original:

{Original Paragraph}
Simplified:

Prompt 2 : B1

Simplify paragraphs for CEFR B1 readers by following these guidelines:
1. **Vocabulary and Sentence Structure:**
 – Use common, everyday words instead of specialized or complex vocabulary.
 – Prefer simple sentence structures; break longer or compound sentences into shorter ones.
 – Replace abstract or complex terms with concrete, clearer expressions or brief explanations.
 – Use active voice where possible and avoid idiomatic expressions or cultural references that may be unclear.
2. **Information Presentation:**
 – Keep all key factual information and core ideas intact to preserve the original meaning.
 – Present numbers, dates, and statistics clearly, often repeating or rephrasing for clarity.
 – When technical or unfamiliar terms appear, define or explain them briefly but simply.
 – Remove less essential details only if they do not affect overall comprehension; otherwise, retain the main content fully.
3. **Clarification and Explicitness:**
 – Make implicit information explicit where needed.
 – Where the original contains pronouns or references that may be unclear, replace or clarify them.
 – Use clear cause–and–effect or chronological connectors (e.g., "because," "so," "however," "since then") to improve coherence.
4. **Tone and Style:**
 – Maintain a neutral, informative, and accessible tone appropriate for learners.
 – Avoid complex or figurative language; use straightforward, literal expressions.
 – When original tone includes subtle nuance, simplify but try to retain the intended emphasis or attitude if important.
5. **Phrasing and Repetition:**
 – Some proper nouns, dates, and well–known names remain unchanged to preserve identity and context.
 – Common phrases and definitions that clarify the subject often get added or slightly expanded to aid understanding.
 – Sentences may be reworded but often echo the original information closely, sometimes repeating key ideas with slight reformulation for clarity.
In summary, simplify vocabulary and grammar, clarify potentially difficult concepts, maintain all essential facts, and ensure readability and coherence for intermediate English learners without omitting important content. Provide only the simplified paragraph without any explanation or justification.

Original:
{Original Paragraph}
Simplified:

One of our four prompts was also generated by an LLM, based on a deductive approach without providing trial data. We made GPT-4.1-mini generate such prompts with the following prompt.

You are an expert in language education and have a deep understanding of CEFR. Your task is to generate a list of rules that an LLM can use to simplify a given English text to each of the CEFR A2 and B1 level. The simplification must

preserve the original meaning and core information of the paragraph. The rules should be specific enough to guide the simplification process effectively, covering vocabulary, grammar, and sentence structure.

As a result, we obtained the following simplification prompts.

Prompt 3 : A2

Simplify paragraphs for CEFR A2 readers by following these guidelines:
Vocabulary
1. Use only high−frequency, common words and phrases (e.g., "big" instead of "enormous," "buy" instead of "purchase").
2. Replace idiomatic expressions and phrasal verbs with their literal or simpler equivalents.
3. Avoid technical or abstract vocabulary; if necessary, define it simply within the sentence.
4. Use concrete nouns and familiar verbs; avoid abstract nouns and complex verbs.
Grammar and Sentence Structure
5. Use simple present and past tenses; avoid complex tenses (e.g., present perfect, future perfect).
6. Favor active voice over passive voice.
7. Use short sentences (ideally 8−12 words per sentence).
8. Use simple sentence structures: one main clause per sentence.
9. Avoid subordinate clauses or replace them with separate sentences.
10. Use coordinating conjunctions (and, but, or) rather than subordinating conjunctions (because, although).
Pronouns and References
11. Limit use of pronouns; repeat nouns when clarity may be lost.
12. Use clear and explicit referents; avoid ambiguous references.
Cohesion and Coherence
13. Use basic discourse markers (first, then, next, finally) to link ideas.
14. Present information in a clear, logical order; avoid complex or backward references.
Miscellaneous
15. Avoid negations where possible; use positive forms or simpler negative constructions.
16. Do not include idioms, humor, or cultural references that might confuse A2 learners.
Provide only the simplified paragraph without any explanation or justification.

Original:
{Original Paragraph}
Simplified:

Prompt 3 : B1

Simplify paragraphs for CEFR B1 readers by following these guidelines:
Vocabulary
1. Use mostly common vocabulary with some intermediate words.
2. Replace very rare or highly technical words with simpler or more familiar synonyms.
3. Allow some idiomatic expressions or phrasal verbs but only if they are common and clearly understandable.
4. Use concrete and abstract nouns where necessary but explain abstract concepts in simple terms.
Grammar and Sentence Structure

5. Use simple present, past, and present perfect tenses; allow future tenses like "will" and "going to."
6. Use both active and passive voice but keep passive structures simple.
7. Use mostly simple sentences and some compound sentences.
8. Allow basic complex sentences with one subordinate clause (e.g., using because, when, if).
9. Avoid very complex sentence structures (e.g., multiple subordinate clauses or embedded clauses).
10. Use modal verbs (can, should, must) appropriately for expressing ability, advice, or obligation.
Pronouns and References
11. Use pronouns more freely than at A2 but ensure clarity is maintained.
12. Use demonstratives (this, that) appropriately for reference.
Cohesion and Coherence
13. Use a wider range of discourse markers (however, although, because, so, therefore) for linking ideas.
14. Present ideas in a logical, well−organized manner with clear paragraphing.
Miscellaneous
15. Include some figurative language if it can be easily explained or inferred.
16. Allow simple conditional sentences (first conditional).
17. Use negations confidently but keep them straightforward (avoid double negatives).
Provide only the simplified paragraph without any explanation or justification.

Original:
{Original Paragraph}
Simplified:

Finally, we prepared one prompt based on basic few-shot (3-shot) prompting. Few-shot examples were extracted from the trial data.

Prompt 4

You are an expert in simplifying complex English texts for language learners. Your task is to adapt the following text to make it accessible and readable for a CEFR {lv} level learner. Provide only the simplified paragraph without any explanation or justification.

Original:
{Original paragraph of the few−shot example 1}
Simplified:
{Reference paragraph of the few−shot example 1}

Original:
{Original paragraph of the few−shot example 2}
Simplified:
{Reference paragraph of the few−shot example 2}

Original:
{Original paragraph of the few−shot example 3}
Simplified:
{Reference paragraph of the few−shot example 3}

Original:
{Original Paragraph}
Simplified:

B Run 2: Prompts for Sentence-level Simplification

B.1 Dynamic Prompt for Sentence-level Simplification (Pseudocode)

```
PROC BUILD_PROMPT(original_sentence,
    influential_phrases,
                   levels={B1,A2,A1},
    add_guardrails=TRUE):

  # 1) Strategy guide (fixed text shown to the
      LLM)
  STRATEGY_GUIDE := [
    "Synonymy (prefer): simpler common words",
    "Modulation: rephrase while preserving
    meaning",
    "Omission: remove non-essential parts",
    "Compression: shorten without losing meaning
    ",
    "Syntactic Changes: simpler structure",
    "Transposition: change word form if clearer",

    "Explanation: only if term appears in
    sentence",
    "Transcription: keep if already simple",
    "Illocutionary Change: make implied meaning
    explicit",
    "DO NOT explain the edits you made"
  ]

  # 2) Level-specific style rules
  RULES.A1 := "1-2 very short sentences; very
    common words only; define technical terms
    briefly if they appear; limited [brackets]."
  RULES.A2 := "1-2 short sentences with simple
    connectors; synonyms for hard words; one
    brief definition at start if needed; at most
     one [bracketed] note."
  RULES.B1 := "Up to two sentences; faithful
    meaning with clearer phrasing; short
    clarifications for technical terms if
    present."

  # 3) Guardrails (optional but recommended)
  IF add_guardrails THEN
    GUARDS := [
      "Only define concepts if the exact word
    appears in the original",
      "Do not add unrelated facts or examples",
      "Preserve names, numbers, and dates
    exactly",
      "Output EXACTLY in the required tag format
    ; no extra commentary"
    ]
  ELSE
    GUARDS := []
  END

  # 4) Select/inject top-K influential phrases
  K := 6
  TOP := TAKE_TOP_BY_ABS_SCORE(
    influential_phrases, K)

  # format for display inside the prompt (one
    per line)
  IF TOP is empty THEN
    INFL_BLOCK := "- (none provided)"
  ELSE
```

```
    INFL_BLOCK := CONCAT_LINES(
      FOR EACH (type, phrase, score) IN TOP:
      "- " + type + " | " + phrase + " : " +
    FORMAT(score, 4 decimals)
    )
  END

  # 5) Assemble prompt text (final structure)
  PROMPT :=
    "You are a CEFR simplification assistant.\n\
    n" +
    "Task:\n" +
    "1) Simplify the original text into these
    CEFR levels: " + JOIN(levels, ", ") + ".\n"
    +
    "2) Focus on these influential/complex
    phrases (apply strategies when needed):\n" +
    INFL_BLOCK + "\n" +
    "3) Keep the meaning accurate but simpler.\n
    " +
    "4) Do not explain what changes you made.\n\
    n" +
    "Strategies:\n" + BULLETIZE(STRATEGY_GUIDE)
    + "\n\n" +
    "Style constraints:\n" +
    "- A1: " + RULES.A1 + "\n" +
    "- A2: " + RULES.A2 + "\n" +
    "- B1: " + RULES.B1 + "\n\n" +
    (IF GUARDS is empty THEN "" ELSE "
    Constraints:\n" + BULLETIZE(GUARDS) + "\n\n"
    ) +
    "Output format (exactly):\n" +
    "<Original> " + original_sentence + "\n" +
    "<B1> ...\n" +
    "<A2> ...\n" +
    "<A1> ..."

  RETURN PROMPT
END
```

B.2 LLM-as-Judge for CEFR Simplification (Pseudocode)

```
PROC BUILD_JUDGE_PROMPT(original_text,
    target_level, candidates, add_guardrails=
    TRUE):

  # 1) Purpose & role
  ROLE := "You are a CEFR simplification judge."

  # 2) Task framing (fixed)
  TASK := [
    "Given: the ORIGINAL sentence, the TARGET
    CEFR level, and several CANDIDATE
    simplifications.",
    "Pick the single BEST candidate that:",
    "1) preserves the meaning of the original,",
    "2) matches the target level (A1 easiest <
    A2 < B1),",
    "3) is fluent and clear,",
    "4) keeps names, numbers, and units correct.
    "
  ]

  \begin{verbatim}
# 3) CEFR decision rubric (concise)
RUBRIC := [
```

```
   "A1: very common words; 1-2 very short
     sentences; concrete, no idioms.",
   "A2: simple frequent words; short sentences;
     simple connectors (and, but); avoid abstract
      terms.",
   "B1: more range; still short clear sentences;
     simple, accurate paraphrase."
]
\end{verbatim}

  # 4) Guardrails (optional)
  IF add_guardrails THEN
    GUARDS := [
      "Return STRICT JSON only (no prose).",
      "Do NOT invent facts; preserve named
    entities, numbers, dates.",
      "Prefer transparent, high-frequency words
    at lower levels.",
      "Minor reordering is fine if meaning is
    preserved.",
      "If ties, choose the clearer and more
    level-appropriate one."
    ]
  ELSE
    GUARDS := []
  END

  # 5) Output contract (exact)
  CONTRACT := [
    "Reply in JSON with:",
    "{ \"winner_index\": <0-based index>, \"
    reason\": \"<one short line>\" }"
  ]

  # 6) Render candidate list
  #    CANDS_STR becomes:
  #    [0] ...
  #    [1] ...
  #    ...
  CANDS_STR := CONCAT_LINES(
    FOR EACH (i, c) IN ENUMERATE(candidates):
      "[" + STR(i) + "] " + c
  )

  # 7) Assemble prompt
  PROMPT :=
    ROLE + "\n\n" +
    JOIN_LINES(TASK) + "\n\n" +
    "CEFR RUBRIC (summary):\n" + BULLETIZE(
    RUBRIC) + "\n\n" +
    (IF GUARDS is empty THEN "" ELSE "
    Constraints:\n" + BULLETIZE(GUARDS) + "\n\n"
    ) +
    JOIN_LINES(CONTRACT) + "\n\n" +
    "ORIGINAL:\n" + original_text + "\n\n" +
    "TARGET LEVEL: " + target_level + "\n\n" +
    "CANDIDATES:\n" + CANDS_STR

  RETURN PROMPT
END
```

C Hyperparameters

In Run 1 (and Run 3), we used HuggingFace Transformers library[3] for open-weight models, namely gpt-oss-40b, Gemma-3-4b, and Qwen-2.5-14b. For

Parameter	Value
temperature	1.0
top_p	0.95
number of beams	1
number of generation	1
max number of new tokens	512

Table 4: Hyperparameters for decoding.

GPT-4.1-mini, we used OpenAI API library.[4] Table 4 shows the hyperparameters settings we used for decoding. For other hyperparameters, default values were used.

D Case Study

In this case study, we analysed one selected example (109-a2) with 51 CEFR-aligned candidates. We carried out a linguistic evaluation of the 19 outputs selected by the filtering process and compared them with the final MBR-selected output. We also examined a set of candidates undervalued by the automatic metrics, some of which provided stronger simplifications than selected outputs.

---

[3]https://huggingface.co/docs/transformers/

[4]https://platform.openai.com/docs/
api-reference/chat/create

Case	Model (Prompt)	Candidate	Sim. Orig.	Sim. Ref.
Orig.	**B2**	Best-selling author Terrie Sharp knows how to tell a good story or two, having twice won the Olive Crime Writer Award. Born on the housing estate with the highest crime rate in Glasgow, she learned to keep quiet, stay out of trouble and how to open car doors with a knife. Luckily, she only used her knowledge to write her novels and she's not so quiet these days, appearing on popular daytime shows like Police Today and at writing conferences throughout the UK.		
Ref.	**A2**	Terrie Sharp is a very popular writer, and her books sell well. She has won the Olive Crime Writer Award two times. She was born in Glasgow, a place with a lot of crime. As a child, she learned to be quiet, stay safe, and even how to open car doors with a knife. Luckily, she only used this for her stories. Now she is not quiet, she goes on TV shows like Police Today and gives talks at writing conferences in the UK.		
MBR	**Qwen-2.5-14b (Prompt 1)**	**Terrie Sharp is a writer who won the Olive Crime Writer Award two times. She was born in a part of Glasgow with lots of crime. She learned to be quiet and avoid trouble. She also learned how to open car doors with a knife, but she only used that for her stories. Now, Terrie speaks a lot on TV shows like Police Today and at events about writing.**	**0.897**	**0.847**
Linguist's note (CEFR A2): This version keeps the meaning of the original while using clear, short sentences. The phrase *avoid trouble* is a simple, transparent way to express the idiom *stay out of trouble*. Using *that* instead of an abstract noun like *knowledge* makes the text easier for A2 readers. The phrase *events about writing* is a little vague compared to *conferences*, but it is easier to understand for learners. Automatic metrics give it a high score mainly because of word overlap, but they do not recognise how the text reduces confusion around the violent detail. Overall, this is a strong fit for A2 learners.			
S1	Gemma (Prompt 4)	Terrie Sharp is a very popular writer. She has won the 'Olive Crime Writer Award' twice! She grew up in an area of Glasgow where there were many problems. She learned to be quiet and not get into trouble. She also learned how to open car doors with a knife. But now, she only uses this knowledge to write her books. She is very busy, and you can see her on TV shows and at writing events in the UK.	0.866	0.845
Linguist's note (CEFR A2): Short sentences help; "problems" is transparent. However, *knowledge* and *appearing/very busy* are abstract for A2; the exclamation skews register. Metrics overestimate due to overlap, not learner appropriateness.			
S2	Gemma (Prompt 4)	Terrie Sharp is a very famous writer. She can write good stories! She has won the Olive Crime Writer Award twice. She grew up in a part of Glasgow where there was a lot of crime. She learned to be quiet and not get into problems. She also learned how to open car doors with a knife. But now, she doesn't use this knowledge anymore. She writes her stories instead. She is on TV shows and talks at writing events in the UK.	0.860	0.843
Linguist's note (CEFR A2): *Famous* is A1/A2 and clearer than *popular*; clause-level simplicity is good. Retains abstract *knowledge*; enthusiastic register ("!") mismatched to neutral bio. Metrics ignore register and reward lexical reuse.			
S3	Gemma (Prompt 2)	Terrie Sharp is a popular writer. She can tell good stories. She won the Olive Crime Writer Award twice. She was born in a part of Glasgow with many crimes. She learned to be quiet and not have problems. She also learned how to open car doors with a knife. She only uses this skill to write her books now. She is not quiet anymore. She appears on TV shows like Police Today. She also goes to writing events in the UK.	0.862	0.861
Linguist's note (CEFR A2): Strong clause splitting; *not have problems* concretizes *stay out of trouble*. But *appears*, *skill* raise lexical load. Metrics laud similarity, not vocabulary transparency.			
S4	Gemma (Prompt 3)	Best-selling writer Terrie Sharp can tell good stories. She won the Olive Crime Writer Award twice. She was born in a part of Glasgow with many crimes. She learned to be quiet and not get into problems. She also learned to open car doors with a knife. She only used this knowledge to write her books. Now, she is not so quiet. She appears on TV shows like Police Today. She also goes to writing events in the UK.	0.860	0.862
Linguist's note (CEFR A2): *Best-selling*, *knowledge* are above A2 transparency; syntax is simple but words remains complex. Metrics overrate due to phrasal reuse.			
S5	Gemma (Prompt 3)	Terrie Sharp is a famous writer. She can tell good stories. She has won the Olive Crime Writer Award twice. She was born in a part of Glasgow with many crimes. She learned to be quiet and not have problems. She also learned how to open car doors with a knife. Now, she doesn't need to use this skill. She is busy. She appears on TV shows like Police Today. She also goes to writing events in the UK.	0.859	0.839
Linguist's note (CEFR A2): This version is clear and easy to follow. Words like *famous* and *busy* are common and simple. The phrase *doesn't need to use* gently reduces the focus on violence, which helps learners. There is a small meaning gap (*busy* is not exactly the same as *not quiet*), but it is acceptable at A2. Automatic metrics reduce the score when the wording changes, even if the change makes the text easier to read.			

Table 5 – continued from previous page

Case	Model (Prompt)	Candidate	Sim. Orig.	Sim. Ref.
S6	Gemma (Prompt 3)	Terrie Sharp is a popular writer. She can tell good stories. She won the Olive Crime Writer Award twice. She grew up in a part of Glasgow with many crimes. She learned to be quiet and not have problems. She also learned how to open car doors with a knife. She uses this knowledge only to write her books. Now, she is not quiet. She appears on TV shows like Police Today. She also goes to writing events in the United Kingdom.	0.859	0.840

Linguist's note (CEFR A2): This version is easy to follow, with a helpful time marker in *grew up*. However, the word *knowledge* is still too abstract for A2 learners, who would benefit more from concrete wording like *what she learned*. Automatic metrics give it credit for overlap, but they do not notice that the vocabulary is still too advanced.

Case	Model (Prompt)	Candidate	Sim. Orig.	Sim. Ref.
S7	Gemma (Prompt 3)	Terrie Sharp is a popular writer. She can tell good stories. She won the Olive Crime Writer Award twice. She was born in a part of Glasgow with many crimes. She learned to be quiet and not have problems. She also learned to open car doors with a knife. She only uses this knowledge to write her books. Now, she is not quiet. She appears on TV shows like Police Today. She also goes to writing events in the UK.	0.860	0.849

Note: Similar to S6: sentence-level clarity, but persistent abstract nouns (*knowledge*, *appears*). Metrics overvalue overlap.

Case	Model (Prompt)	Candidate	Sim. Orig.	Sim. Ref.
S8	GPT-4.1-mini (Prompt 4)	Terrie Sharp is a famous writer who knows how to tell good stories. She won the Olive Crime Writer Award two times. She grew up in a poor area of Glasgow where there was a lot of crime. There, she learned to be quiet, avoid trouble, and how to open car doors with a knife. But she only used these skills for her books. Now, she talks on TV shows like Police Today and visits writing events in the UK.	0.856	0.768

Linguist's note (CEFR A2): *Poor area* is an easy-to-understand paraphrase of *housing estate with the highest crime rate* (it adds a slight suggestion of poverty). The phrases *avoid trouble* and *talks/visits* are clear and simple. Automatic metrics often mark down synonymy even when the result is a better fit for A2 learners.

Case	Model (Prompt)	Candidate	Sim. Orig.	Sim. Ref.
S9	GPT-4.1-mini (Prompt 2)	Terrie Sharp is a famous writer who knows how to tell stories. She has won the Olive Crime Writer Award two times. She was born in a part of Glasgow with a lot of crime. There, she learned to be quiet and avoid trouble. She also learned how to open car doors with a knife. She used this knowledge only to write her books. Now, she talks a lot on TV shows like Police Today. She also goes to writing events in the UK.	0.854	0.782

Linguist's note (CEFR A2): Clear, concrete phrasing; *talks a lot* makes the contrast explicit. *Knowledge* remains abstract. Automatic metrics often do not give enough credit for phrasal simplification.

Case	Model (Prompt)	Candidate	Sim. Orig.	Sim. Ref.
S10	GPT-4.1-mini (Prompt 3)	Terrie Sharp is a writer who writes good stories. She won the Olive Crime Writer Award two times. She was born in a part of Glasgow with many crimes. She learned to be quiet and safe. She also learned how to open car doors with a knife. She only used this skill for her books. Now, she talks on TV shows like Police Today. She also goes to writing meetings in the UK.	0.863	0.866

Linguist's note (CEFR A2): The phrase *writer who writes* is repetitive. On the positive side, *safe* is a good simplification. Automatic metrics reward word overlap, but they ignore problems of fluency and style.

Case	Model (Prompt)	Candidate	Sim. Orig.	Sim. Ref.
S11	GPT-4.1-mini (Prompt 3)	Terrie Sharp is a very good writer. She won the Olive Crime Writer Award two times. She was born in a part of Glasgow where many crimes happen. She learned to be quiet and to avoid trouble. She also learned how to open car doors with a knife. She did not use this skill for bad things. She uses it to write books. Now, she talks on TV shows like Police Today. She also goes to writing meetings in the UK.	0.855	0.850

Linguist's note (CEFR A2): The line *not for bad things* adds a clear moral point, which reduces ambiguity and helps A2 readers. The wording is concrete throughout. Automatic metrics usually don't give enough credit for these helpful additions.

Case	Model (Prompt)	Candidate	Sim. Orig.	Sim. Ref.
S12	GPT-4.1-mini (Prompt 3)	Terrie Sharp is a writer. She won the Olive Crime Writer Award two times. She was born in a part of Glasgow with a lot of crime. She learned to be quiet and stay safe. She also learned how to open car doors with a knife. She only used this skill to write her books. Now, she talks on TV shows like Police Today. She also goes to writing meetings in the UK.	0.855	0.812

Linguist's note (CEFR A2): This version is concise and easy to read. Some details may be lost, but it still matches A2 aims (short clauses, simple connectors). Automatic metrics often mark down compression even when the result remains adequate for learners.

Case	Model (Prompt)	Candidate	Sim. Orig.	Sim. Ref.
S13	Qwen-2.5-14b (Prompt 4)	Terrie Sharp is a writer who sells lots of books. She has won an award two times for writing stories about crimes. She was born in a part of Glasgow with many crimes. She learned to be quiet and not get in trouble. She even learned how to open car doors with a knife! But she never used that skill in real life. Instead, she writes stories about them. Now, she talks on TV shows like Police Today and at big writer meetings all over the UK.	0.868	0.865

Table 5 – continued from previous page

Case	Model (Prompt)	Candidate	Sim. Orig.	Sim. Ref.
Linguist's note (CEFR A2): Easy words, but some facts are not correct (*award for writing about crimes*) and the tone is too dramatic by adding("!").				
S14	Qwen-2.5-14b (Prompt 4)	Terrie Sharp, a famous writer, knows how to tell stories really well. She has won an award twice for writing about crimes. She grew up in a place in Glasgow with a lot of crime. She learned to be quiet and not get in trouble, and even how to open car doors with a knife. But now she uses those stories to write her books instead. Today, Terrie is not so quiet anymore. She goes on TV shows like Police Today and speaks at many writing events in the UK.	0.859	0.874
Linguist's note (CEFR A2): The text is simple and easy to read, but it adds details that are not in the original (*award for writing about crimes*; *uses those stories*). These changes make the text less accurate. Automatic metrics give a high score for word overlap, but they fail to notice the meaning errors.				
S15	Qwen-2.5-14b (Prompt 4)	Terrie Sharp is a popular writer who won an award twice. She was born in a place in Glasgow where there were many crimes. She learned to stay quiet and not get in trouble. She even learned how to open car doors with a knife! But she didn't use those skills for bad things. Instead, she wrote stories about them. Now, she talks a lot and appears on TV shows like Police Today and at writing events all over the UK.	0.856	0.857
Linguist's note (CEFR A2): Clear explanation of right and wrong; small loss of detail (award name) and a small added idea ("wrote stories about them"). Scores accept these changes but do not notice omissions.				
S16	Qwen-2.5-14b (Prompt 2)	Terrie Sharp is a best-selling author who writes crime stories. She has won the Olive Crime Writer Award twice. She was born in a part of Glasgow with a high crime rate. There, she learned to be quiet and not get in trouble. She used to know how to open car doors with a knife but only used that for her books. Now, Terrie is more outgoing. She appears on TV shows like Police Today and goes to writing events across the UK.	0.858	0.820
Linguist's note (CEFR A2): Adds extra detail (crime stories) and keeps a hard phrase ("high crime rate"). "Outgoing" changes the meaning of "quiet." Scores go down a bit for these changes, not for word difficulty.				
S17	Qwen-2.5-14b (Prompt 3)	Terrie Sharp is a writer who won the Olive Crime Writer Award twice. She was born in a neighborhood in Glasgow with a lot of crime. She learned to be quiet and avoid trouble. She also learned how to open car doors with a knife. But she only used this for her stories. Now, Terrie talks a lot on TV shows like Police Today. She also goes to writing meetings all over the UK.	0.856	0.846
Linguist's note (CEFR A2): The word *neighborhood* is an easy and familiar synonym that works well for A2 learners. Using *this* instead of an abstract noun makes the meaning clearer. The phrase *talks a lot* shows the contrast directly. This is a strong A2 version.				
S18	Qwen-2.5-14b (Prompt 3)	Terrie Sharp writes good stories. She won the Olive Crime Writer Award twice. She was born in a dangerous part of Glasgow. There, she learned to stay quiet and avoid trouble. She even learned to open car doors with a knife. But she never used this skill for bad things. Now, she uses her stories to appear on popular TV shows like Police Today. She also speaks at writing events across the UK.	0.866	0.842
Linguist's note (CEFR A2): *Dangerous part* is an excellent simplification. But causal inversion (*uses her stories to appear*) is a semantic error. Metrics miss meaning coherence.				
S19	Sentence-lv	Terrie Sharp is a good storyteller. She won the Olive Crime Writer Award twice. She was born in a place in Glasgow with a lot of crime. She learned to be silent, avoid problems, and open car doors with a knife. She writes novels using her knowledge. Now, she talks on quiet TV programs in the UK, like Police Today and writing seminars.	0.875	0.723
Linguist's note (CEFR A2): *Silent* is a over-literal change for *quiet*; *quiet TV programs* is a meaning error; *seminars* and *knowledge* are hard words for A2. The low score shows meaning changes, but metrics cannot pinpoint error type.				

Table 5: Example 109-a2 from the test data. The table presents the original B2 paragraph, the human reference simplification at A2, the final MBR-selected output from Run 3, and 19 candidates selected by the filtering system. Rows shaded red are judged (very good), rows shaded green are judged (excellent) and rows shaded yellow are *poor/problematic* and need revision. Among all candidates, **S8 (GPT-4.1-mini, Prompt 4)** is the best candidate for CEFR A2 pedagogical, balancing lexical simplicity, syntactic ease, semantic fidelity, and learner support through explicit moral clarification.

Case	Model/Prompt	Selection (Non-Selected)	Orig. Sim.	Ref. Sim.
NS1	Gemma-3-4b (Prompt 4)	Terrie Sharp is a very popular writer of stories. She has won the Olive Crime Writer Award two times. She grew up in a place in Glasgow with many crimes. She learned to be quiet and not get into trouble. She also learned how to open car doors with a knife. But she used this only to write her books. Now she is not quiet. She goes on TV shows like Police Today and to writing events in the UK.	0.828	0.802

Linguist's note: This version uses short, clear sentences. The phrase "writer of stories" avoids the business-like idea of "books sell well." It keeps the violent detail but explains it as only used for writing. Automatic metrics mark it down because the wording differs from the reference, but it is still very suitable for A2.

Case	Model/Prompt	Selection (Non-Selected)	Orig. Sim.	Ref. Sim.
NS2	Gemma-3-4b (Prompt 4)	Terrie Sharp is a famous writer who can write many stories. She won the Olive Crime Writer Award two times. She was born in a part of Glasgow with many crimes. She learned to be quiet and not have problems. She also learned how to open car doors with a knife. But now she uses it only for her books. Today she talks a lot on TV shows like Police Today and at writing events in the UK.	0.848	0.797

Linguist's note : The word "famous" is clearer than "popular," and "not have problems" is a good replacement for the idiom "stay out of trouble." The sentences are simple and easy to follow. The only weakness is a small redundancy ("can write many stories"). Metrics lower the score because of different words, not because of difficulty.

Case	Model/Prompt	Selection (Non-Selected)	Orig. Sim.	Ref. Sim.
NS3	Gemma-3-4b (Prompt 1)	Terrie Sharp is a popular writer. She has won the Olive Crime Writer Award twice. She was born in a part of Glasgow with many crimes. She learned to be quiet and not have problems. She also learned how to open car doors with a knife. Now she only uses this for her stories. Today she talks on TV shows like Police Today and at writing events in the UK.	0.830	0.830

Linguist's note: The sentences are clear and short. The idiom "stay out of trouble" is replaced with the simpler "not have problems." All key facts are kept. Metrics give only average credit, even though this is strong A2 material.

Case	Model/Prompt	Selection (Non-Selected)	Orig. Sim.	Ref. Sim.
NS4	Gemma-3-4b (Prompt 1)	Terrie Sharp is a popular writer. She has won the Olive Crime Writer Award twice. She was born in a part of Glasgow with many crimes. She learned to be quiet and not have problems. She also learned how to open car doors with a knife. Now she only uses this for her stories. She talks on TV shows like Police Today and goes to writing events in the UK.	0.848	0.762

Linguist's note: This version has the same strengths as NS3. The similarity score is low because it says "events" instead of "conferences," but "events" is actually easier for A2 learners.

Case	Model/Prompt	Selection (Non-Selected)	Orig. Sim.	Ref. Sim.
NS5	Gemma-3-4b (Protmp 1)	Terrie Sharp is a famous writer. She writes good stories. She has won the Olive Crime Writer Award two times. She was born in a part of Glasgow with many crimes. She learned to be quiet and not have problems. She also learned how to open car doors with a knife. She now uses this only to write her books. She goes on TV shows like Police Today and to writing events in the UK.	0.853	0.816

Linguist's note: The words "famous" and "writes good stories" are simple and transparent. The sentences are split clearly. The meaning is faithful to the original. Metrics reduce the score only because the wording does not match the reference.

Case	Model/Prompt	Selection (Non-Selected)	Orig. Sim.	Ref. Sim.
NS6	Gemma-3-4b (Prompt 2)	Terrie Sharp is a popular writer. She can tell good stories. She won the Olive Crime Writer Award twice. She was born in a part of Glasgow with many crimes. She learned to be quiet and not have problems. She also learned how to open car doors with a knife. She only uses this for her books now. She is not quiet anymore. She talks on TV shows like Police Today and goes to writing events in the UK.	0.854	0.804

Linguist's note: The time marker "now" makes the story easier to follow. The contrast "not quiet anymore" is clear. The only issue is the abstract noun "knowledge," which remains a bit hard for A2. Metrics again don't give full credit for these improvements.

Case	Model/Prompt	Selection (Non-Selected)	Orig. Sim.	Ref. Sim.
NS7	Gemma-3-4b (Prompt 2)	Terrie Sharp is a popular writer. She can tell good stories. She won the Olive Crime Writer Award twice. She was born in a part of Glasgow with many crimes. She learned to be quiet and not have problems. She also learned how to open car doors with a knife. She uses this only to write her books now. She is not quiet anymore. She talks on TV shows like Police Today and goes to writing events in the UK.	0.849	0.780

Linguist's note: Very similar to NS6: short sentences, easy connectors, clear contrast. The similarity score is lower because of harmless word changes.

Case	Model/Prompt	Selection (Non-Selected)	Orig. Sim.	Ref. Sim.
NS8	Gemma-3-4b (Prompt 3)	Terrie Sharp is a famous writer. She can tell good stories. She has won the Olive Crime Writer Award twice. She was born in a part of Glasgow with many crimes. She learned to be quiet and not have problems. She also learned how to open car doors with a knife. She only uses this to write her books now. She goes on TV shows like Police Today and to writing events in the UK.	0.852	0.786

Linguist's note: The verbs "goes on" and "uses" are clear and concrete. Sequencing is simple and effective. Metrics score it lower because the words differ from the reference.

Table 6 – continued from previous page

Case	Model (Prompt)	Candidate (Non-Selected)	Sim. Orig.	Sim. Ref.
NS9	GPT-4.1-mini (Prompt 4)	Terrie Sharp is a famous writer. She has won the Olive Crime Writer Award two times. She was born in Glasgow, a place with a lot of crime. She learned to be quiet and not get into trouble. She also learned to open car doors with a knife. She only used this to write her books. Now, she talks on TV shows like Police Today and goes to writing meetings in the UK.	0.828	0.834
Linguist's note: This version is very clear and accurate. Words like *famous* and *meetings* are easy for A2 learners. The sequencing is simple and the contrast is clear. Automatic metrics give it a lower score because of different wording, but it is an excellent A2 simplification.				
NS10	GPT-4.1-mini (Prompt 1)	Terrie Sharp is a famous writer who has won the Olive Crime Writer Award two times. She was born in a part of Glasgow with a lot of crime. She learned to be quiet and avoid trouble. She also learned how to open car doors with a knife. She used this only to write her books. Now, she talks on TV shows like Police Today and goes to writing meetings in the UK.	0.842	0.770
Linguist's note: This version handles the idiom well (*avoid trouble*), and the word "meetings" is culturally simple at A2. The content is faithful and the style is easy to read. The lower similarity score only reflects useful word changes, not quality loss.				

Table 6: Case study analysis of **non-selected** outputs that were linguistically strong but scored lower on automatic metrics. Rows shaded red are judged (very good) and rows shaded green are judged (excellent). These examples show that metrics often mark down simplifications that use common words (e.g., *famous* vs. *popular*, *meetings* vs. *conferences*) and concrete phrasing, even though they better match CEFR A2 descriptors.
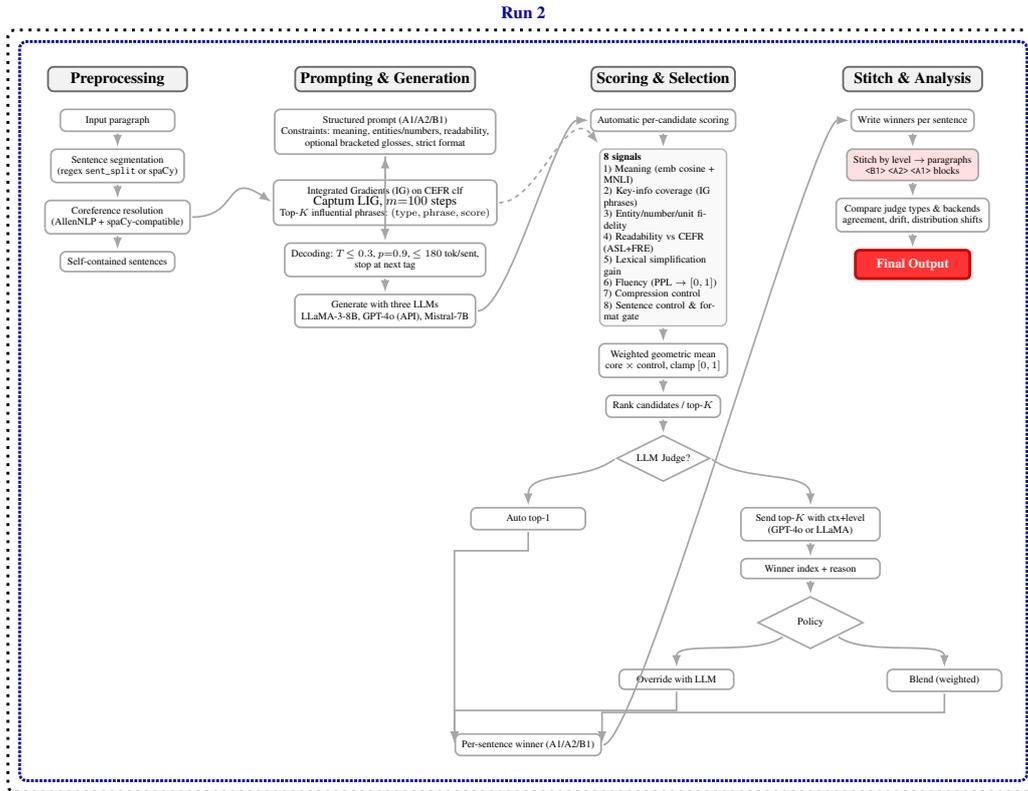
E Sentence Simplification architecture: Run2

Figure 3: Run 2 preprocessing (segmentation, coreference) and IG attribution; CEFR-controlled prompting/decoding across three LLMs; automatic judge (8 signals) with weighted geometric mean; optional LLM-as-Judge with policy; stitching and comparative analysis.

F Evaluation-Metrics-Sentence simplification: Run2

Metric	Description
Meaning preservation	Embedding cosine similarity plus bidirectional entailment probabilities (MNLI) to assess whether the simplified sentence preserves the meaning of the source.
Key information coverage	Checks whether the top-K influential phrases identified by IG are present in the simplified output (case-insensitive matching).
Entity, number, and unit fidelity	Compares named entities with spaCy (set F1). Numbers are greedily matched one-to-one if units agree, allowing an absolute error within $\max(1\%, 10^{-6})$.
Readability vs. CEFR	Combines average sentence length (ASL) and Flesch Reading Ease (FRE) (Flesch, 1948), normalised to CEFR targets: A1 (ASL $\approx$ 10, FRE $\geq$ 0.80), A2 (15, 0.70), B1 (20, 0.60).
Lexical simplification gain	Reduction in average syllables per word compared to the source. A small bonus is given for inline glosses (e.g., "[*a simple meaning*]").
Fluency	Language model perplexity mapped to $[0, 1]$ (Jurafsky and Martin, 2023); lower perplexity means higher fluency. If no LM is provided, a neutral score of 0.75 is assigned.
Compression control	Ratio of simplified to original word counts, normalised to the target range 0.6–1.0. Penalises outputs that are too short or too verbose.
Sentence/format control	Encourages keeping sentence count close to the source (ratio 0.7–1.1). Rejects empty outputs or those exceeding 1200 characters.

Table 7: Evaluation signals used by the automatic judge. Each metric is normalised to $[0, 1]$ and combined by a weighted geometric mean.