Text simplification in a low-resource setting: Benchmark study on Finnish

Anonymous ACL submission

Abstract

Text simplification aims to make text easier to read while preserving meaning, with applications in critical areas, such as accessibility and second-language education. Although it has been extensively studied for English, lower-resource languages remain under-explored. We present the first benchmark for Finnish text simplification with explicit focus on CEFR-based difficulty control—a pedagogically motivated approach essential for creating graded reading materials. To address the problem of scarcity of parallel datasets, we synthesize a training set of 1,123 pairs using LLM-based generation with difficulty filtering. We manually curate a high-quality dataset of 143 samples from over 5,000 aligned pairs, applying strict criteria for difficulty reduction and semantic preservation. Systematic evaluation of nine models—including 2 fine-tuned Finnish LLMs, 2 base models, and 5 commercial systems—reveals that fine-tuned Ahma achieves superior CEFR-level difficulty reduction (1.25 CEFR level on average) with distinctive pedagogical strategies, significantly outperforming ChatGPT (0.53 CEFR level). However, human evaluation reveals that commercial models like Gemini produce more fluent, error-free text, demonstrating superior overall quality despite comparable difficulty reduction scores.

1 Introduction

Text simplification has been an active research area in natural language processing (NLP) for several decades. Its goal is to modify text to make it easier to read and understand without loss of semantic content. Numerous applications for text simplification include improving accessibility, enhancing communication, and supporting language education. While many studies have focused on English text simplification, simplification in lower-resource languages is still under-explored.

In this paper, we focus on simplification of Finnish text with specific emphasis on *difficulty reduction* rather than generic simplification. While recent work on text simplification has explored neural approaches (Sheang and Saggion, 2024) and multilingual evaluation (Saggion et al., 2022; Dmitrieva and Tiedemann, 2024), these efforts typically optimize for general readability without explicit difficulty targeting. In contrast, building on recent advances in CEFR-based difficulty assessment (Wu et al., 2026), our work explicitly targets controlled CEFR-level reduction—a pedagogically motivated approach essential for creating graded reading materials. Our work explicitly targets reductions in linguistic difficulty as measured on the CEFR scale, while preserving semantic content. This focus is crucial for educational applications, where learners require graded reading materials adapted to specific proficiency levels.

One major challenge in our study, as in prior work, is the scarcity of parallel data for training and evaluation. To address this issue, we distill knowledge from a readability-controlled text generation pipeline, similar to that of Katinskaia et al. (2025), judged by an accurate Finnish difficulty estimation model (Wu et al., 2026). We use this pipeline to create a synthetic parallel dataset for Finnish text simplification, train specialized models, and evaluate their performance on a manually curated test set. Our evaluation measures not only semantic preservation and fluency, but critically assess whether models can produce text at lower CEFR levels than source material.

Our main contributions are:

- A new **dataset** for Finnish text simplification with CEFR-based difficulty control, including a synthetic dataset (1,123 pairs) and a manually curated set (143 samples selected from 5,000+ aligned pairs).

- The first **benchmark** for Finnish simplification, with a comprehensive evaluation of

nine models across automatic metrics and human evaluation. The experiments reveal that fine-tuned models significantly outperform their training data source (Ahma-SFT: 1.25 vs. ChatGPT: 0.53 points).

- We **manually evaluate** the results of simplification from multiple viewpoints, and identify important differences between fine-tuned models and commercial models, such as the ability to exhibit unique pedagogical strategies.

The paper is organized as follows. Section 2 reviews related work. In Section 3, we describe the data used for evaluation. In Section 4, we discuss our methodology. In Section 5, we present the experiments and results. In Section 6, we present the human evaluation. We conclude in Section 7.

2 Related Work

Most prior work on simplification focuses on English. Zhu et al. (2010), Zhang and Lapata (2017) and Coster and Kauchak (2011) create parallel datasets based on English Wikipedia. This inspired the creation of other datasets (Xu et al., 2016). Xu et al. (2015) introduced the Newsela dataset, which is manually annotated by professional editors. Moving beyond English text simplification, Gala et al. (2020) released Alector, a dataset for simplification in French. Dmitrieva and Tiedemann (2021) and Dmitrieva et al. (2021) created parallel datasets based on Russian novels and their simplified versions. These works have provided resources for developing and evaluating text simplification models (Guo et al., 2018; Martin et al., 2022; Jiang et al., 2020; Omelianchuk et al., 2021).

Although there are a number of datasets in English, few datasets have finer-grain annotations using a *readability scale* (Uchida et al., 2018; Xu et al., 2015; Scarton et al., 2018). Xu et al. (2015) and Scarton et al. (2018) follow the Common Core Standards (Porter et al., 2011), while Uchida et al. (2018) follows the Common European Framework of Reference for Languages (CEFR). In our work, we follow the CEFR standard as well.

The pipeline for simplification in Finnish of Katinskaia et al. (2025) operates as follows: ChatGPT is prompted to simplify each text to a specified target difficulty level, and a critic model evaluates whether the difficulty has been reduced. If the reduction is insufficient, the process iterates until a satisfactory simplification is achieved. For this study, we replace the original critic model with the Finnish difficulty estimation model of Wu et al. (2026). This model was evaluated against expert CEFR annotations and achieves an error close to the level of inter-annotator disagreement among human experts (MAE of 8.24% on the CEFR difficulty scale), indicating that its predictions are nearly as reliable as expert human judgment. We therefore adopt it as the difficulty estimator for both data filtering and evaluation throughout this work. While the iterative simplification pipeline provides a baseline for text simplification, our experiments reveal significant limitations that motivate the need for a dedicated simplification model trained on Finnish data. Specifically, we observe that ChatGPT-based simplification achieves poor results for Finnish texts: the system frequently fails to reduce difficulty at all, and when simplification does occur, it is often minimal and unpredictable. Furthermore, the iterative nature of the pipeline—requiring multiple API calls with a critic model in the loop—results in high computational costs and long processing times, making it impractical for large-scale application or real-time use.

CEFR Level	Difficulty Score	CEFR Level	Difficulty Score
A1	1	B2	4
A1-A2	1.5	B2-C1	4.5
A2	2	C1	5
A2-B1	2.5	C1-C2	5.5
B1	3	C2	6
B1-B2	3.5		

Table 1: Mapping from CEFR levels to difficulty scores

3 Data

3.1 Training Data

The challenges observed with the baseline simplification for Finnish underscore the need for a specialized model that can perform reliable, controllable simplification efficiently. To this end, we generate training data using the baseline pipeline described in Section 2 as a bootstrapping mechanism, despite its limitations. We begin by extracting texts at difficulty levels B1 and above—B1, B2, C1, and C2—from a Finnish difficulty assessment dataset (Wu et al., 2026).

We set multiple target difficulty levels to cover two primary simplification scenarios: C→B transformations (with targets C1-C2, B2-C1, and B2 all reducing toward B1) and B→A transformations

(with targets B1-B2, B1, and A2 all reducing toward A2). Note that A2→A1 targets were treated as A2→A2, since A1 is too simplistic to be reliably achieved by automated methods; we therefore only require observable difficulty scores reduction for these very few cases.

When used in the pipeline (Katinskaia et al., 2025), ChatGPT proved inadequate for Finnish text simplification. In many cases, the model fails to reduce the difficulty at all, and when simplification does occur, it often results in smaller reductions than intended. For quality control, we retain only those generated pairs where the difficulty score decreased by at least 0.5 points—i.e., at least *half* of a CEFR level. Table 1 shows the mapping between CEFR levels and difficulty scores used in our study. The most common successful simplifications were single-level reductions such as C2→C1 or B2→B1, rather than the broader multi-level reductions originally targeted.

Table 2 summarizes the training data statistics. For texts originally at difficulty level B1 or B2, the average reduction in difficulty was 0.68 points, yielding only 94 usable pairs. For texts at level C1 or C2, we achieved better results with an average reduction of 1.59 points, obtaining 1,029 pairs from around 2,000 documents. These modest results further illustrate the limitations of the LLM-based pipeline, and highlight the need for a model specifically trained for Finnish simplification. Note that in Table 2, the "Original Level" represents the actual CEFR level of the source text, while the target level (B→A and C→B) represent the *intended* simplification scenario rather than the achieved outcome, as discussed above. Further details appear in Appendix B.

3.2 Test Data

We derive our test data from the Language Bank of Finland (Kielipankki),[1] a repository of Finnish textual corpora, including parallel corpora for text simplification[2] (Dmitrieva and Tiedemann, 2024). The data consists of aligned sentence pairs from original and simplified Finnish texts. It is important to note that while the corpus provides alignment between original and simplified versions, only a very small portion of the aligned pairs represent genuine text simplification—some pairs are identical, and many exhibit only superficial modifications without meaningful reduction in difficulty.

[1] Language Bank of Finland, Aineistot

[2] Corpus LB-2024011701, Corpus LB-2022111625

Simplification Type	Avg. Reduction (levels / points)	Pairs	%
C→B	1.59	1,029	91.6%
B→A	0.68	94	8.4%
Total	**1.53**	**1,123**	**100%**

Table 2: Training data statistics

To construct a high-quality evaluation set, we manually curated suitable parallel samples according to strict criteria. First, we require *semantic preservation*: the simplified version must neither omit essential information from the source nor introduce new information not present in the source. Second, we require *measurable reduction in the difficulty* level between the original and simplified versions. To ensure reliability, we employed a two-stage verification process: pairs were first manually inspected for semantic fidelity, then validated using the difficulty estimation model (Wu et al., 2026) to confirm that the simplified text indeed exhibits lower difficulty than the original.

This curation process yields 143 high-quality parallel sample pairs from over 5,000 aligned pairs in the original Kielipankki corpus. Of these, 92 samples were reserved as our test set, while the remaining 51 samples were used to augment the training set described in Section 3.1.

We consider the simplified text in each pair in this constructed parallel corpus as the *ground truth*. Hereafter, we refer to the parallel simplified Finnish test corpus as *Selko* (i.e., "simplified Finnish"), which serves as the test set for all subsequent evaluations. More details about this dataset are provided in Appendix B.

3.3 Data Quality

As shown in Table 2, the substantial difference in sample sizes between C-level and B-level sources—1,029 vs. 94 pairs—reflects the inherent challenges of the simplification task at different difficulty levels. ChatGPT demonstrates greater success at simplifying advanced (C-level) texts, possibly because these texts contain more complex linguistic features that can be targeted for simplification. In contrast, B-level texts, being closer to intermediate proficiency, may already exhibit relatively straightforward language, leaving less room for substantial simplification while preserving semantic content.

We evaluate all models on our manually curated test set derived from authentic Selko corpus. This provides a more robust assessment of real-world applicability, as the test set reflects genuine human

simplification rather than ChatGPT's generation patterns.

4 Methodology

4.1 Model Training

We perform supervised fine-tuning (SFT) on Poro 2.0 (Zosa et al., 2025) and Ahma-7B,[3] two LLMs specifically developed for Finnish and other Nordic languages. Fine-tuning was performed using training data described in Section 3.1, with hyperparameters presented in Appendix A, Table 11.

The combination of a small batch size per-device (1) with 16 gradient accumulation steps yields an effective batch size of 16, which balances memory constraints with stable training dynamics. The relatively low learning rate (5e-6) was chosen to prevent catastrophic forgetting of the base model's Finnish language capabilities, while allowing for adaptation to the simplification task.

4.2 Experimental Setup

We compare a diverse set of models to establish a comprehensive benchmark for Finnish text simplification, spanning across several categories:

Proprietary Large Language Models:

- **ChatGPT**—GPT-5.4 (OpenAI, 2023); this was also used to generate the base of our training data.
- **Gemini**—Google's latest multimodal model (Reid et al., 2024). We use Gemini 3 Flash.
- **DeepSeek**—open-weights model with strong multilingual capability (DeepSeek-AI, 2024). We use DeepSeek v4 flash in this paper.
- **Grok-2**—conversational model (xAI, 2024). We use grok-4.3 in this paper.
- **Qwen**—Alibaba's instruction-tuned model (Team, 2024). We use qwen-3.6-flash.

Finnish-Specialized Base Models:

- **Poro**—Base Finnish language model without fine-tuning. We used Poro 2.0 in this paper.
- **Ahma**—Base Finnish language model without fine-tuning.

Models fine-tuned on our training data:

- **Poro-SFT**—Poro 2.0 fine-tuned.
- **Ahma-SFT**—Ahma fine-tuned.

This allows us to evaluate and compare: (1) the effectiveness of general-purpose LLMs on low-resource Finnish simplification, (2) the baseline performance of Finnish-specialized models, and (3) the improvement gained through supervised fine-tuning on our synthetic training data.

[3] Ahma-7B

Metric	Description
MAE	Mean absolute error
RMSE	Root mean square error
Acc@0.5	Accuracy within 0.5 difficulty points
Acc@1.0	Accuracy within 1.0 difficulty point
Pearson	Linear correlation coefficient
Spearman	Monotonic correlation coefficient

Table 3: Difficulty reduction metrics

4.3 Evaluation Metrics

Our evaluation framework encompasses three complementary dimensions: difficulty reduction, semantic preservation, and surface-level similarity.

Difficulty Reduction: We evaluate the effectiveness of difficulty reduction using automated difficulty estimation model (Wu et al., 2026), which assigns numerical difficulty scores as described in Table 1. For each test sample, we compute the difficulty scores of three texts: the original complex text, the model-generated simplified text, and the ground-truth simplified text.

Our evaluation uses six metrics (Table 3). The first four metrics assess difficulty reduction. Pearson's r and Spearman's ρ measure correlation with reference texts: human simplifications—to evaluate alignment with gold standards, and original complex texts—to verify difficulty preservation.

Semantic Similarity: To assess how well simplified texts preserve the original meaning, we use cosine similarity between sentence embeddings.[4] We compute four semantic similarity metrics:

- Average similarity (of simplified text) to original text.
- Average similarity (of simplified text) to ground truth.
- High Similarity Rate: Proportion with similarity between simplified and original text > 0.8.
- Low Similarity Rate: Proportion with similarity between simplified and original text < 0.5.

These metrics provide nuanced insights into how well the models respect semantic fidelity.

[4] https://huggingface.co/TurkuNLP/sbert-cased-finnish-paraphrase

Metric	Description
corpus_BLEU	Corpus-level BLEU score
avg_sentence_BLEU	Average sentence-level BLEU
BLEU-1	Unigram precision
BLEU-2	Bigram precision
BLEU-3	Trigram precision
BLEU-4	4-gram precision

Table 4: BLEU metrics for surface-level similarity

Simplification Quality: SARI To assess the quality of simplification operations, we employ SARI (System output Against References and against the Input sentence) (Xu et al., 2016). Unlike BLEU, which measures only surface similarity to the reference, SARI measures three operations: (1) adding helpful n-grams, (2) retaining necessary n-grams, and (3) deleting complex n-grams.

The final SARI score (ranging from 0 to 100) is the arithmetic mean of these three sub-scores. SARI has become the standard automatic metric for text simplification evaluation (Alva-Manchego et al., 2020), as it better captures simplification-specific operations than translation-oriented metrics. While SARI was originally designed for English, its n-gram-based approach is language-agnostic and has been successfully applied to morphologically rich languages (Dmitrieva and Tiedemann, 2024). For Finnish specifically, the metric can accommodate the language's agglutinative morphology by treating inflected forms as distinct, thus capturing both lexical and morphological simplification. In this paper, we present 4 metrics related to SARI:

- SARI: Overall SARI score (0-100).
- SARI std: Standard deviation of SARI scores.
- High SARI Rate: Proportion with SARI > 40.
- Low SARI Rate: Proportion with SARI < 30.

Surface-level Similarity: BLEU For completeness and comparability to other work, we also report BLEU scores comparing model outputs to ground-truth simplifications. We note that BLEU has significant limitations for evaluating text simplification, particularly for morphologically rich languages like Finnish. As noted by Sulem et al. (2018), BLEU was designed for machine translation and rewards overlap of surface-form n-grams, which does not reflect the validity of simplification operations, such as lexical substitution, sentence splitting, or paraphrasing—operations that may use entirely different word forms while preserving meaning. This limitation is exacerbated in Finnish due to agglutinative morphology, where a single lemma can generate dozens of inflected forms depending on case, number, possession, etc. (Pirinen, 2015). Therefore, we primarily rely on SARI for assessing simplification quality, using BLEU only as a supplementary reference.

5 Experiments and Results

5.1 Performance on Reduction of Difficulty

Table 5 presents the difficulty reduction results comparing model outputs against the Selko ground truth. Ahma-SFT achieves the best performance with an average reduction of 1.25 CEFR points (which means more than one CEFR level) and a 98.9% reduction rate. Gemini and Qwen also perform strongly, achieving reduction of 1.22 and 1.18, respectively. In terms of accuracy metrics (Acc@0.5 and Acc@1.0), which measure how closely outputs match the target difficulty, Qwen performs comparably to Ahma-SFT. Notably, ChatGPT—the model we used to generate our training data—achieves only 0.53 points of reduction, on average. This demonstrates that supervised fine-tuning on imperfect synthetic data outperforms the general-purpose model that generated the data. Among Finnish-specialized models, Poro-SFT (0.94 points) lags behind Ahma-SFT (1.25 points) despite identical fine-tuning. We hypothesize that this gap stems from their origin: Poro was adapted from English Llama, whereas Ahma was trained from scratch on Finnish corpora, likely giving it a stronger intrinsic understanding of Finnish linguistic structure.

Table 6 compares model outputs against the original (complex) texts. Ahma-SFT, Gemini, and Qwen lead with MAE of 1.19–1.26, demonstrating substantial difficulty reduction. Besides, Ahma-SFT achieves the lowest Acc@0.5 (0.10), meaning only 10% of outputs stay within 0.5 points of the original—evidence of more effective simplification. In contrast, ChatGPT shows the strongest correlation with original texts (Pearson: 0.83) and highest Acc@1.0 (0.84), revealing a conservative strategy that maintains proximity to source difficulty rather than pursuing aggressive simplification.

5.2 Semantic Preservation

Table 7 presents the scores on semantic similarity between model outputs and reference texts. Commercial LLMs demonstrate exceptional semantic preservation, with all models achieving average

Model	Avg. Reduction	Reduction Rate (%)	MAE	RMSE	Acc @0.5	Acc @1.0	Pearson r	Spearman ρ
ChatGPT	0.53	87.0	0.75	0.88	0.38	0.65	0.77	0.70
Gemini	1.22	**98.9**	**0.36**	**0.46**	0.68	**0.97**	0.79	0.67
Deepseek	0.72	92.4	0.68	0.85	0.43	0.75	0.72	0.60
Grok	1.05	96.7	0.43	0.57	0.70	0.92	0.77	0.67
Qwen	1.18	96.7	0.39	0.52	**0.73**	0.92	0.76	0.73
Poro	0.77	89.1	0.68	0.83	0.39	0.82	0.42	0.42
Ahma	0.81	84.8	0.70	0.86	0.40	0.75	0.45	0.41
Poro-SFT	0.94	93.5	0.57	0.73	0.53	0.86	0.55	0.51
Ahma-SFT	**1.25**	**98.9**	0.39	0.50	0.70	0.93	0.72	0.69

Table 5: Difficulty reduction performance compared to Selko (ground truth)

Model	MAE	RMSE	Acc @0.5	Acc @1.0	Pearson r	Spearman ρ
ChatGPT	0.55	0.69	0.51	0.84	0.83	0.78
Gemini	1.22	1.35	0.12	**0.35**	0.70	0.60
Deepseek	0.76	0.93	0.39	0.67	0.77	0.71
Grok	1.06	1.20	0.18	0.50	0.75	0.68
Qwen	1.19	1.34	0.17	0.36	0.66	0.62
Poro	0.87	1.04	0.32	0.55	0.48	0.46
Ahma	0.94	1.10	0.29	0.57	0.51	0.49
Poro-SFT	1.02	1.20	0.22	0.53	0.50	0.51
Ahma-SFT	**1.26**	**1.39**	**0.10**	**0.35**	0.65	0.62

Table 6: Difficulty reduction performance vs. original complex texts

similarity above 0.93 to original texts and above 0.88 to Selko references. Gemini and Grok achieve the highest similarity to Selko (0.89), with over 91% of their outputs showing high similarity (>0.8). Notably, no commercial model produced outputs with low similarity (<0.5), indicating robust semantic preservation during simplification.

Among the models specialized for Finnish, fine-tuning significantly improves semantic preservation. Ahma-SFT outperforms base Ahma, with the high-similarity rate increasing from 64.1% to 89.1%. Poro-SFT (0.845) also improves over base Poro (0.825). Both Finnish models still lag behind commercial LLMs in semantic fidelity. Notably, Poro is the only model that produced *any* low-similarity outputs, suggesting occasional semantic drift without fine-tuning. These results demonstrate that while all approaches successfully maintain core semantic content, commercial models and fine-tuned Finnish models achieve more consistent preservation across all test samples.

5.3 SARI Scores

Table 8 presents SARI scores measuring simplification *quality*. Gemini achieves the highest score (44.05), followed by strong performers including Ahma-SFT (42.36) and Grok (42.60). We note that Ahma-SFT significantly outperforms ChatGPT (39.63)—the model used to generate its training data—demonstrating that supervised fine-tuning successfully learns superior simplification strategies beyond what the *teacher* model can produce through zero-shot prompting. Poro-SFT (40.13) also surpasses base Poro (39.21), confirming the effectiveness of fine-tuning. Among all non-commercial models, Ahma-SFT achieves a 57.6% high-quality rate (SARI > 40) with only 1.1% low-quality outputs.

5.4 BLEU Scores

Table 9 reports BLEU scores for comparison with prior work. Ahma-SFT achieves the highest corpus BLEU of 23.14, substantially exceeding all other models and demonstrating strong surface-level alignment with reference simplifications. Among commercial models, Grok (13.57) and Deepseek (9.60) show the next-highest BLEU scores. Notably, Ahma-SFT's BLEU-1 through BLEU-4 scores (75.0, 34.8, 13.6, 9.5) surpass those of base Ahma (38.5, 5.3, 1.4, 0.7), indicating that fine-tuning substantially improves lexical overlap with human references.

6 Human Evaluation

To complement the automatic metrics, we conducted a human evaluation with a native Finnish speaker. We randomly sampled 12 test instances and their simplifications from all nine models. The

Model	Avg vs. Original	Avg vs. Selko	High Similarity Rate (> 0.8)	Low Similarity Rate (< 0.5)
ChatGPT	0.945	0.878	89.1%	0.0%
Gemini	0.934	**0.890**	**93.5%**	0.0%
Deepseek	**0.968**	0.885	90.2%	0.0%
Grok	0.950	**0.890**	91.3%	0.0%
Qwen	0.946	0.888	91.3%	0.0%
Poro	0.861	0.825	75.0%	1.1%
Ahma	0.837	0.807	64.1%	0.0%
Poro-SFT	0.868	0.845	78.3%	0.0%
Ahma-SFT	0.893	0.870	89.1%	0.0%

Table 7: Semantic similarity evaluation results

Model	SARI	SARI Std	High SARI Rate (> 40)	Low SARI Rate (< 30)
ChatGPT	39.63	3.99	45.7%	2.2%
Gemini	**44.05**	5.40	**78.3%**	**0.0%**
Deepseek	42.09	4.75	64.1%	**0.0%**
Grok	42.60	4.39	69.6%	**0.0%**
Qwen	42.34	5.27	63.0%	**0.0%**
Poro	39.21	7.07	38.0%	5.4%
Ahma	38.24	6.51	28.3%	9.8%
Poro-SFT	40.13	7.27	42.4%	4.3%
Ahma-SFT	42.36	**7.72**	57.6%	1.1%

Table 8: SARI scores for text simplification quality. Number of valid samples is 92

evaluator assessed each output across three dimensions: (1) *difficulty reduction*, whether the text became easier to understand—rated as Strong, Moderate, or Weak; (2) *semantic preservation*, whether the core meaning was maintained—rated as Strong, Moderate, or Weak; and (3) *textual errors*, the presence of grammatical mistakes, factual inaccuracies, or unnatural expressions. Given the scale of this evaluation (12 samples × 9 models × 3 dimensions) and space constraints, the main text focuses on a direct comparison between the strongest open-source model (Ahma-SFT) and the strongest commercial model (Gemini), which are the most informative contrast for our research question. The complete human evaluation results for all nine models are provided in the supplementary material.

6.1 Difficulty Reduction

Table 10 presents human judgments on readability. Since our evaluator lacked formal CEFR annotation training, we followed Martinc et al. (2021) and asked the evaluator to judge whether each text showed an improvement in *readability*. Gemini was rated Strong in 92% of cases (11/12), whereas Ahma-SFT received Strong ratings in only 25% (3/12), with the rest rated Moderate. This contrasts sharply with our automatic metrics (Table 5), where the two models achieved nearly identical average difficulty reductions (1.25 vs. 1.22 levels). Consistent with Vajjala and Lučić (2018), this shows that measurable CEFR-level reduction does not always align with perceived readability: human readers are also sensitive to fluency and cohesion, dimensions that automatic difficulty metrics do not capture.

6.2 Semantic Preservation

Table 10 shows human ratings for semantic preservation. The evaluator judges whether the simplified text retained the meaning and key information from the original, without adding content. Gemini received mostly Strong ratings (92%, 11/12), while Ahma-SFT received only 17% Strong ratings (2/12), despite achieving comparable or even higher automatic semantic similarity scores in several cases. For example, Ahma-SFT outputs with similarity scores of 0.957 and 0.954 were rated Moderate, while Gemini outputs with lower scores (0.914) received Strong ratings.

This discrepancy between automatic metrics and human judgment can be attributed to the fact that Ahma-SFT occasionally introduces factual inaccuracies (Section 6.3), such as numerical distortions or misattributed information. These factors significantly impact the perceived semantic preservation, despite high embedding similarity.

Model	Corpus BLEU	Avg Sent. BLEU	BLEU-1	BLEU-2	BLEU-3	BLEU-4
ChatGPT	9.24	7.66	63.2	22.2	5.9	3.1
Gemini	8.12	10.10	**75.0**	20.0	7.1	3.8
Deepseek	9.60	**10.48**	66.7	28.6	**15.4**	4.2
Grok	13.57	9.98	73.9	27.3	9.5	2.5
Qwen	7.51	9.14	66.7	21.4	7.7	4.2
Poro	5.36	3.48	45.5	9.5	2.5	1.3
Ahma	3.53	3.47	38.5	5.3	1.4	0.7
Poro-SFT	6.61	3.42	60.0	21.1	2.8	1.5
Ahma-SFT	**23.14**	6.55	**75.0**	**34.8**	13.6	**9.5**

Table 9: BLEU scores across different models

Dimension	Model	Automatic	Human Rating
Difficulty	Gemini	1.22	**92% Strong** 08% Moderate
	Ahma-SFT	**1.25**	25% Strong 75% Moderate
Semantic	Gemini	**0.914–0.986**	**92% Strong** 08% Moderate
	Ahma-SFT	0.799–0.961	17% Strong 83% Moderate

Table 10: Automatic metrics vs. human judgments (on 12 samples)

6.3 Error Analysis

The evaluator agreed that Gemini produced error-free outputs across all 12 samples. In contrast, Ahma-SFT exhibited factual and grammatical errors in several instances. As shown in the example below, the original text (top) reports specific measurements: Sodankylä's Vuotso had 97 cm of snow, while Salla's Naruska had 96 cm—only 1 cm less. The Ahma-SFT output (bottom) introduces obvious factual distortions:

> **Original:** *"Eniten lunta koko maassa on Sodankylässä. Ilmatieteen laitoksen mukaan Sodankylän Vuotsossa on 97 senttiä lunta. Vain sentin vähemmän lunta eli 96 senttiä on Sallan Naruskalla."* (The most snow in the entire country is in Sodankylä. According to the Finnish Meteorological Institute, there is 97 centimeters of snow in Vuotso, Sodankylä. Only one centimeter less snow, or 96 centimeters, is in Naruska, Salla.)
>
> **Ahma-SFT:** *"Ilmatieteen laitos kertoi, että Lapissa satoi viikonloppuna yli 80 senttimetriä lunta, mutta vain yksi sentti enemmän satoi Rovaniemelle."* (The Finnish Meteorological Institute reported that over 80 centimeters of snow fell in Lapland over the weekend, but only one centimeter more fell in Rovaniemi.)

The original 97 cm figure refers to *accumulated* snow depth in Vuotso, not "weekend" snowfall, and the "1 cm difference" describes the comparison between Vuotso (97 cm) and Naruska (96 cm) in the original, not Rovaniemi. This demonstrates that Ahma-SFT can conflate or misinterpret numerical information during simplification, a failure not observed in Gemini outputs.

6.4 Discussion

The evaluator identified three distinctive capabilities unique to Ahma-SFT that were not observed in other models: (1) systematic term definition—consistently introducing *explanations* for technical terms like "kaamos" and "Eurovision"; (2) syntactic decomposition—breaking complex participial phrases into simpler clauses with transparent connectives; and (3) semantic redundancy elimination—merging duplicate information at the propositional level rather than surface paraphrase. These patterns align with principles of pedagogical simplification and suggest genuine learned simplification *strategies*. However, factual accuracy errors, particularly numerical distortions, remain a limitation that automatic metrics fail to capture, underscoring the need for higher-quality training data. The full evaluator analysis is provided in Appendix E.

7 Conclusion

This work demonstrates that supervised fine-tuning on limited synthetic data substantially improves Finnish text simplification. Ahma-SFT achieved 1.25 CEFR points reduction, and significantly outperformed ChatGPT (0.53 points), despite being trained on ChatGPT-generated data. However, important limitations remain: models cannot reliably control the degree of reduction to target specific CEFR levels, and Ahma-SFT can inject factual errors, absent in commercial models like Gemini. In future work we plan to address these issues by (1) constructing larger corpora using Gemini; (2) exploring reinforcement learning for fine-grained difficulty control; and (3) developing multi-stage simplification pipelines to enable precise targeting for diverse reader populations.

Limitations

We should note the several limitations in this work in its present state:

- **Data Scarcity**: The training dataset is limited in size (1,123 pairs), owing to the labor-intensive process of collecting and curating high-quality parallel simplification data, especially for a lower-resourced language such as Finnish. We consider this a critical limitation and leave its mitigation to future work.

- **Dependency on ChatGPT**: Our synthetic training data is generated by ChatGPT. Any change in ChatGPT's behavior or accessibility may affect the reproducibility of our pipeline. Moreover, since the target side of the training pairs is ChatGPT output, our fine-tuned models partly distill ChatGPT's simplification behavior; their gains on automatic metrics should be read in this light.

- **Scale of Human Evaluation**: Our human evaluation was conducted by a single native Finnish speaker on 12 test instances. We were therefore unable to compute inter-annotator agreement, and the results should be regarded as a preliminary, qualitative complement to the automatic metrics rather than a definitive quantitative assessment. The detailed analysis in the main text focuses on the strongest open-source and commercial models; complete results for all nine models are provided in the supplementary material. A larger-scale evaluation with multiple trained annotators is left for future work.

References

Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. 2020. ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4668–4679, Online. Association for Computational Linguistics.

William Coster and David Kauchak. 2011. Simple English Wikipedia: A new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 665–669, Portland, Oregon, USA. Association for Computational Linguistics.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Anna Dmitrieva, Antonina Laposhina, and Maria Lebedeva. 2021. A quantitative study of simplification strategies in adapted texts for L2 learners of russian. In *Proceedings of the International Conference "Dialogue*, pages 191–203.

Anna Dmitrieva and Jörg Tiedemann. 2021. Creating an aligned Russian text simplification dataset from language learner data. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 73–79, Kiyv, Ukraine. Association for Computational Linguistics.

Anna Dmitrieva and Jörg Tiedemann. 2024. Towards automatic Finnish text simplification. In *Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024*, pages 39–50, Torino, Italia. ELRA and ICCL.

Núria Gala, Anaïs Tack, Ludivine Javourey-Drevet, Thomas François, and Johannes C. Ziegler. 2020. Alector: A parallel corpus of simplified French texts with alignments of misreadings by poor and dyslexic readers. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1353–1361, Marseille, France. European Language Resources Association.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2018. Dynamic multi-level multi-task learning for sentence simplification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 462–476, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. 2020. Neural CRF model for sentence alignment in text simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7943–7960, Online. Association for Computational Linguistics.

Anisia Katinskaia, Anh-Duc Vu, Jue Hou, Ulla Vanhatalo, Yiheng Wu, and Roman Yangarber. 2025. Estimation of text difficulty in the context of language learning. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 594–611, Vienna, Austria. Association for Computational Linguistics.

Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. 2022. MUSS: Multilingual unsupervised sentence simplification by mining paraphrases. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1651–1664, Marseille, France. European Language Resources Association.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2021. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, 47(1):141–179.

Kostiantyn Omelianchuk, Vipul Raheja, and Oleksandr Skurzhanskyi. 2021. Text Simplification by Tagging. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 11–25, Online. Association for Computational Linguistics.

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

Tommi A Pirinen. 2015. Omorfi — free and open source morphological lexical database for Finnish. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 313–315, Vilnius, Lithuania. Linköping University Electronic Press, Sweden.

Andrew Porter, Jennifer McMaken, Jun Hwang, and Rui Yang. 2011. Common core standards: The new U.S. intended curriculum. *Educational Researcher*, 40:103–116.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Horacio Saggion, Sanja Štajner, Daniel Ferrés, Kim Cheng Sheang, Matthew Shardlow, Kai North, and Marcos Zampieri. 2022. Findings of the TSAR-2022 shared task on multilingual lexical simplification. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 271–283, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Carolina Scarton, Gustavo Paetzold, and Lucia Specia. 2018. Text simplification from professionally produced corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Kha Chau Sheang and Horacio Saggion. 2024. Controllable text simplification with lexical constraint loss. In *Proceedings of ACL 2024*, pages 1891–1905.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. BLEU is not suitable for the evaluation of text simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 738–744, Brussels, Belgium. Association for Computational Linguistics.

Qwen Team. 2024. Qwen2.5: A party of foundation models. *arXiv preprint arXiv:2412.15115*.

Satoru Uchida, Shohei Takada, and Yuki Arase. 2018. CEFR-based lexical simplification dataset. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Sowmya Vajjala and Ivana Lučić. 2018. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 297–304, New Orleans, Louisiana. Association for Computational Linguistics.

Yiheng Wu, Jue Hou, and Roman Yangarber. 2026. Translation as augmentation: Effect of translated data on assessment of difficulty. In *Proceedings of the 2nd Workshop on Evaluating Text Difficulty in a Multilingual Context (DeTermIt!) LREC 2026*.

xAI. 2024. Grok-1. Accessed: 2024-12-01.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 584–594, Copenhagen, Denmark. Association for Computational Linguistics.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1353–1361, Beijing, China. Coling 2010 Organizing Committee.

Elaine Zosa, Jouni Louma, Kai Hakala, Antti Virtanen, Mika Koistinen, Risto Luukkonen, Akseli Reunamo, Sampo Pyysalo, and Jonathan Burdge. 2025. Poro 2: Continued pretraining for language acquisition. LumiOpen.

A Appendix: Fine-tuning setup

Table 11 shows the values of hyper-parameters used for fine-tuning the Poro 2 (7b parameters) and Ahma (7b parameters) model.

B Details about Training and Test Data

Table 12, Table 13, Table 14 and Table 15 in this appendix show the details of the breakdown of the instances in the training and test datasets. The instances in these datasets are *pairs*: original source sentence → simplified sentence—and this simplification was done by ChatGPT. The selection of instances for fine-tuning and evaluation of our simplification model based on Poro is described in Table 2.

Hyperparameter	Value
Per-device train batch size	1
Gradient accumulation steps	16
Learning rate	5.0×10^{-6}
Max gradient norm	1.0
Warmup ratio	0.01
Number of training epochs	10
rank	8

Table 11: Fine-tuning hyperparameters

We draw the "original" sentences from the dataset for difficulty assessment in Finnish, as mentioned in Section 3.

Original → Simplified	Count
C1-C2 → B1-B2	235
C1-C2 → B2	229
C1-C2 → B2-C1	140
C1-C2 → B1	98
C1-C2 → C1	61
C1 → B2	56
C1 → B1	49
C1 → B1-B2	39
C1 → B2-C1	20
C2 → B1-B2	19
C2 → C1	17
C2 → C1-C2	17
C2 → B2-C1	12
C2 → B2	10
C1 → A2-B1	10
C1-C2 → A2-B1	8
C2 → B1	5
C1-C2 → A2	2
C2 → A2	1
C1 → A2	1
Total	**1,029**

Table 12: C→B simplification: we ask ChatGPT to simplify these instances from their original level C to B1. The Table shows the counts of the outcomes of these simplifications.

C LLM prompt

In this paper, we used the same prompts uniformly for all LLMs to simplify text. We used two prompts corresponding to the simplification of two difficulty levels: C2->B2 and B2->A2.

For the C2->B2 level, we used the following prompt:

You are an expert in the Finnish language and language teaching. You will be given a text in Finnish. Your task is to read it first and then to provide an adaptation into CEFR level B1. Do not significantly change the meaning of the input text, please. B1 is an intermediate level. The definition of a learner with B1 level is: "Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can produce simple connected text on topics, which are familiar, or of personal interest. Can describe experiences and events, dreams, hopes & ambitions and briefly give reasons and explanations for opinions and plans. Imagine that you are teaching a B1 learner, your adaptation should fit their proficiency level. This is the text to simplify: text

For the B2->A2 level, we used the following prompt:

You are an expert in the Finnish language and language teaching. You will be given a text in Finnish. Your task is to read it first and then to provide an adaptation into CEFR level A2. Do not significantly change the meaning of the input text, please. A2 is just above the beginner level. Texts in A2 should be simple, with relatively easy grammar. Imagine that you are teaching someone, who just recently started to learn a language. Your adaptation should fit their proficiency level. This is the text to simplify: text'

D Summary of human evaluation

Here we only have list statistics on human evaluation from both the Ahma-SFT and Gemini models. We attach the full human evaluaion results in supplyment material.

Original → Simplified	Count
B1-B2 → B1	47
B2 → B1-B2	16
B1-B2 → A2-B1	15
B1-B2 → A2	6
B1 → A2-B1	5
B2 → B1	2
B2 → A2-B1	2
B1 → A2	1
Total	**94**

Table 13: B→A simplification: we ask ChatGPT to simplify these instances from their original level B to A2. The Table shows the counts of the outcomes of these simplifications

Original → Simplified	Count
C1-C2 → B2	6
B2-C1 → B2	5
C1-C2 → B2-C1	5
C1 → B1-B2	4
B1 → A1-A2	3
B2-C1 → B1-B2	3
C1-C2 → B1-B2	3
C1-C2 → B1	3
A2 → A1-A2	2
B1-B2 → A2	2
B2 → B1-B2	2
B1 → A2-B1	1
B1-B2 → A2-B1	1
B2 → B1	1
B2 → B2	1
B2 → A2-B1	1
B2-C1 → A2-B1	1
B2-C1 → B2-C1	1
C1 → A2-B1	1
C1 → B2	1
C1 → B1	1
C1 → C1	1
C1 → B2-C1	1
Total	**51**

Table 14: Pairs of original and manually simplified sentences from Kielipankki, which we manually curated and selected as instances for our *training* data.

Original → Simplified	Count
C1-C2 → B2	22
B2-C1 → B1-B2	11
C1-C2 → B2-C1	8
C1-C2 → B1-B2	8
C1 → B1-B2	6
C1 → B2	4
C1-C2 → C1	4
B2 → B1	3
B2 → B2	3
B2-C1 → B2	3
B1 → A1-A2	2
B2 → B1-B2	2
B2 → A2	2
B2-C1 → B1	2
B2-C1 → B2-C1	2
C1 → B1	2
C1-C2 → B1	2
A2 → A1-A2	1
B1 → A2	1
B1-B2 → B1-B2	1
B1-B2 → B1	1
B2 → A2-B1	1
B2-C1 → A2-B1	1
Total	**92**

Table 15: Pairs of original and manually simplified sentences from Kielipankki, which we manually curated and selected as instances for our *test* data.

E Qualitative analysis by human evaluator

E.1 Systematic Introduction of Term Definitions

Ahma-SFT consistently introduces a *definition* when a term first appears in the text—unlike the other models, for which this is the exception rather than the rule. This is a standard practice in manual production of simplified text, and Ahma-SFT applies it systematically across the entire dataset.

Example 1: Eurovision

> **Input:** *Tämän vuoden euroviisujen voittaja on Ruotsi, jolle Suomi antoi 12 pistettä.* (This year's Eurovision winner is Sweden, to which Finland gave a point.)
>
> **Ahma-SFT:** *Euroviisut ovat kilpailu, jossa valitaan Euroopan paras laulu. Tänä vuonna kilpailun voitti Ruotsi.* (Eurovision is a competition to choose the best song from Europe. This year, Sweden won the competition.)

ChatGPT, DeepSeek, Gemini and Grok move straight to the results without explaining what Eurovision is.

Example 2: Kaamos

> **Input:** *Kaamos alkoi tiistaina Suomen pohjoisimmassa kylässä Utsjoen Nuorgamissa.* (The storm began on Tuesday in Finland's northernmost village, Nuorgam, Utsjoki.)
>
> **Ahma-SFT:** *Kaamos tarkoittaa aikaa talvella, kun aurinko ei näy taivaalla. Kaamos kestää melkein kaksi kuukautta.* (Kaamos refers to the time in winter when the sun is not visible in the sky. Kaamos lasts for almost two months.)

ChatGPT, Gemini and Grok do not explain the term at all.

	Ahma-SFT	Gemini
Strong Difficulty Reduction	3/12 (25%)	11/12 (92%)
Moderate Difficulty Reduction	9/12 (75%)	1/12 (8%)
Strong Semantic Preservation	2/12 (17%)	11/12 (92%)
Moderate Semantic Preservation	10/12 (83%)	1/12 (8%)
Factual Errors	3	0
Syntax Errors	0	0
Logic Errors	0	0
Total Errors	3	0

Table 16: Summary of human evaluation: Ahma-SFT vs. Gemini (12 samples)

Example 3: ICAN

> **Input:** *Rauhanpalkinto annetaan tänä vuonna ydinaseita vastaan työskentelevälle Ican-verkostolle.* (This year's Peace Prize will be awarded to the Ican network, which works against nuclear weapons.)
>
> **Ahma-SFT:** *Palkinto annetaan ihmisille, jotka taistelevat ydinaseita vastaan. Se annettiin Ican-nimiselle järjestölle.* (The award is given to people who fight against nuclear weapons. It was given to an organization called Ican.)

The organisation is first described through its function, and only then named—something ChatGPT, DeepSeek and Grok do not do.

E.2 Breaking Down Syntactically Complex Constructions

Ahma-SFT more frequently than other models reduces syntactically overloaded constructions to simpler units—replacing participial phrases and genitive chains with either two simple sentences or a subordinate clause with a more transparent connective.

Example 1: Participial phrase decomposition

> **Input:** *Rauhanpalkinto annetaan tänä vuonna ydinaseita vastaan työskentelevälle Ican-verkostolle.*(This year's Peace Prize will be awarded to the Ican network, which works against nuclear weapons.)
>
> **Ahma-SFT:** *Palkinto annetaan ihmisille, jotka taistelevat ydinaseita vastaan. Se annettiin Ican-nimiselle järjestölle.* (The award is given to people who fight against nuclear weapons. It was given to an organization called Ican.)

The participial phrase is split into two sentences with a *jotka*-clause. DeepSeek and ChatGPT leave the construction intact.

Example 2: Genitive chain simplification

> **Input:** *Manilan Ninoy Aquinon kansainvälinen lentokenttä on suljettu tulivuorenpurkauksen aiheuttaman tuhkapilven johdosta.* (Manila's Ninoy Aquino International Airport has been closed due to an ash cloud caused by a volcanic eruption.)
>
> **Ahma-SFT:** *Lentokenttä kaupungin lähistöllä on suljettu, koska purkaus aiheuttaa ilmassa tuhkan, kiven ja kaasun sekoitusta, joka lentää jopa yli 60 kilometriä tunnissa.* (The airport near the city has been closed because the eruption is sending a mixture of ash, rock and gas into the air, flying at speeds of up to 60 kilometers per hour.)

The chain of infinitives and genitive attributes is replaced by a causal clause with *koska*, which is easier to read for a beginner.

E.3 Eliminating Semantic Redundancy

In several cases Ahma-SFT identifies and removes meaning-level repetitions present in the source text, which other models reproduce faithfully. This suggests the model operates to some degree at the level of propositional content rather than surface-level paraphrase.

Example 1: Double mention elimination

> **Input:** *Tämän vuoden euroviisujen voittaja on Ruotsi, jolle Suomi antoi 12 pistettä. Ruotsi on voittanut tämän vuoden*

euroviisut. (The winner of this year's Eurovision Song Contest is Sweden, to which Finland gave 12 points. Sweden has won this year's Eurovision Song Contest.)

Ahma-SFT: *Tänä vuonna kilpailun voitti Ruotsi.* (This year, the competition was won by Sweden.)

The winner is named once. DeepSeek and Grok repeat *voitti* twice within three consecutive sentences.

Example 2: Headline-body duplication removal

Input (headline): *Palkinnon saa ydinaseiden vastainen Ican-järjestö*(The prize goes to the anti-nuclear organization Ican)

Input (body): *Rauhanpalkinto annetaan tänä vuonna ydinaseita vastaan työskentelevälle Ican-verkostolle* (This year's Peace Prize will be awarded to the Ican network working against nuclear weapons.)

Ahma-SFT: *Nobelin rauhanpalkinto myönnetään tänään. Palkinto annetaan ihmisille, jotka taistelevat ydinaseita vastaan. Se annettiin Ican-nimiselle järjestölle.* (The Nobel Peace Prize is being awarded today. The prize is given to people who fight against nuclear weapons. It was given to an organization called Ican.)

Both propositions are merged into a single coherent statement without repetition.

F Implementation Details for Evaluation Metrics

SARI: We used the implementation from the Hugging Face datasets library (version 3.2.0).
BLEU: We employed the BLEU implementation from the datasets library (version 3.2.0).
Sentence Transformer: We utilized the sentence-transformers library (version 3.3.1) for computing semantic similarity.