Let's Simplify Step by Step: Guiding LLM Towards Multilingual Unsupervised Proficiency-Controlled Sentence Simplification

Jingshen Zhang[1,2], Xin Ying Qiu[1]*, Lifang Lu[1], Zhuhua Huang[1],
Yutao Hu[1], Yuechang Wu[1], JunYu Lu[3]

[1]Department of Computer Science, School of Information Science and Technology,
Guangdong University of Foreign Studies, Guangzhou, China
[2]College of Intelligence and Computing, Tianjin University, Tianjin, China
[3]Lionrock AI Lab, China Merchants Research Institute of Advanced Technology, China
audbut0702@163.com, xy.qiu@foxmail.com

Abstract

Large language models demonstrate limited capability in proficiency-controlled sentence simplification, particularly when simplifying across large readability levels. We propose a framework that decomposes complex simplifications into manageable steps through dynamic path planning, semantic-aware exemplar selection, and chain-of-thought generation with conversation history for coherent reasoning. Evaluation on five languages across two benchmarks shows our approach improves simplification effectiveness while reducing computational steps by 22-42%. Human evaluation confirms the fundamental trade-off between simplification effectiveness and meaning preservation. Notably, even human annotators struggle to agree on semantic preservation judgments, highlighting the inherent complexity of this task. Our work shows that while step-by-step simplification improves control, preserving semantic fidelity during extensive simplification remains an open challenge.

1 Introduction

Controlled text simplification, introduced by Scarton and Specia (2018), aims to automatically rewrite sentences to match the comprehension level of a specific target audience. This capability is essential for creating educational materials for children (Javourey-Drevet et al., 2022), adapting information for broader audiences (Grabar and Cardon, 2018), and ensuring content accessibility across languages (Ryan et al., 2023).

The field of controllable text simplificationen compasses diverse control mechanisms and methodological approaches. Control strategies can be categorized as either binary distinctions between complex and simple text (Blinova et al., 2023) or ordinal scales based on specific criteria such as age of acquisition (Oshika et al., 2024), simplicity rank (Paetzold and Specia, 2015; Glavaš and Štajner, 2015; Qiang et al., 2021), school grade level (Scarton and Specia, 2018; Agrawal and Carpuat, 2023; Cripwell et al., 2023; Mo and Hu, 2024), readability scores (Imperial and Madabushi, 2023; Farajidizaji et al., 2024), or CEFR proficiency levels (Imperial and Madabushi, 2023; Barayan et al., 2025). The scope of simplification varies across granularities, ranging from document-level (Cripwell et al., 2023; Blinova et al., 2023; Mo and Hu, 2024) and paragraph-level (Imperial and Madabushi, 2023; Farajidizaji et al., 2024) to sentence-level (Scarton and Specia, 2018; Agrawal and Carpuat, 2023; Barayan et al., 2025) and lexical-level transformations (Glavaš and Štajner, 2015; Paetzold and Specia, 2015; Qiang et al., 2021; Oshika et al., 2024). Methodologically, the field has evolved from early approaches based on machine translation and machine learning (Paetzold and Specia, 2015; Scarton and Specia, 2018) to pre-trained language model (PLM) based methods (Qiang et al., 2021; Cripwell et al., 2023; Blinova et al., 2023; Agrawal and Carpuat, 2023) and, more recently, large language model (LLM) based approaches (Imperial and Madabushi, 2023; Farajidizaji et al., 2024; Mo and Hu, 2024; Oshika et al., 2024; Barayan et al., 2025). The present study focuses specifically on CEFR proficiency-controlled sentence simplification using LLM-based methods.

CEFR proficiency-level sentence simplification represents controllable simplification aligned with the Common European Framework of Reference for Languages, a widely recognized framework for describing language ability across six levels (A1 to C2). Advances in this area include the development of specialized corpora such as the CEFR-based Sentence Profile (CEFR-SP) corpus (Arase et al., 2022) and multilingual datasets such as README++ (Naous et al., 2024) and UniversalCEFR (Imperial et al., 2025).

Recent studies have revealed notable limitations

---

*Corresponding author: xy.qiu@foxmail.com

in LLMs' ability to perform proficiency controlled complexity adjustments. For example, Barayan et al., 2025 demonstrated success in moderate simplification tasks. However, LLMs often struggle with more challenging scenarios that require larger readability spans, such as simplifying from C2 to A1 level. Additionally, our pilot studies revealed a trade-off between reaching the target simplification level and preserving semantic meaning.

To address these challenges, we propose decomposing complex simplification into manageable incremental steps. Building on Chain-of-Thought reasoning (Wei et al., 2023), we develop a novel framework that combines systematic path planning with semantic-aware exemplar selection. We evaluate this approach through comprehensive multilingual experiments across 5 languages with extensive human assessment. Our code is available at: https://github.com/JasonZhang0702/simplify-step-by-step

The key contributions of our work are the following:

A dynamic programming approach for optimal path planning that decomposes large-span simplifications into intermediate steps, improving both target level achievement (by up to 20 percentage points) and computational efficiency (22-42% reduction in inference steps)

A semantic-aware Chain-of-Thought framework that combines exemplar selection based on meaning preservation with conversation history tracking, enabling coherent multi-step simplification while maintaining semantic fidelity.

Comprehensive multilingual evaluation across 5 languages with extensive human assessment, revealing a fundamental trade-off between readability control and meaning preservation that intensifies with simplification span—a challenge confirmed by both automatic metrics and human experts.

2 Related Research

Controllable text simplification: Controllable text simplification aims to adapt text complexity to specific target audiences or proficiency levels. As noted in our introduction, control objectives vary from binary classifications to fine-grained ordinal scales. Controllable approaches include explicit simplification operations (Alva-Manchego et al., 2017; Dong et al., 2019), attribute based control (Martin et al., 2020), paraphrasing (Maddela et al., 2021) or control tokens prepended to input sequences to pre-trained language models (Agrawal and Carpuat, 2023).

For CEFR-controlled generation specifically, Stowe et al. (2022) explored controlled generation using a concept2seq framework with CEFR and semantic role labels as control features, though they limited generation to only A1 or C2 extremes. Imperial and Madabushi (2023) examined proficiency alignment of instruction-tuned LLMs for passage-level simplification, finding that more information in prompts yields better results. Barayan et al. (2025) analyzed CEFR-controlled sentence simplification across eight LLMs, demonstrating that models struggle with lower readability targets. Our paper particularly addresses the challenges of simplifying across multiple proficiency levels, where direct single-step simplification may result in insufficient simplification.

Planning-Based Approaches in Text Generation: Planning mechanisms have emerged as effective strategies for improving controlled text generation across various tasks. For text summarization, hierarchical planning decomposes generation into content selection and surface realization stages (Moryossef et al., 2019). Puduppully et al. (2019) proposed a neural data-to-text generation model that explicitly incorporates content selection and planning stages. For text simplification specifically, Cripwell et al. (2023) proposed a document planning approach for document-level simplification, where an LLM first generates a high-level plan outlining the document structure before performing sentence-level simplifications. They showed that planning significantly outperforms end-to-end document simplification models. Related work on sentence deletion (Zhong et al., 2020; Zhang et al., 2022) uses discourse structure features from surrounding sentences to identify deletion candidates. In our work, we apply dynamic path planning to determine optimal intermediate proficiency levels between source and target, for more efficient controlled in simplification across large span of levels.

Few-Shot Chain-of-Thought Generation Few-shot learning with carefully selected exemplars has proven effective for text generation tasks. Liu et al. (2022) showed that selecting in-context examples (relative to random sampling) that are semantically-similar to a test sample improves generation quality for Large Language Model. Malik et al.(2024) demonstrated that selecting exemplars based on minimizing control error improves target level achievement in CEFR-controlled

generation. While we control simplification accuracy with path-planning, our semantic-aware exemplar selection method prompts LLMs with best sentence pairs from validation stage for improved meaning preservation in simplification generation.

Chain-of-Thought (CoT) prompting (Wei et al., 2023) has emerged as a powerful technique for eliciting step-by-step reasoning in LLMs across various tasks. Kojima et al.(2022) showed that even zero-shot CoT with simple prompts like "Let's think step by step" can improve reasoning. Zhou et al.(2023) introduced least-to-most prompting with a top-down decomposition of the problem and a bottom-up resolution generation. The idea is to break down a complex problem into a series of simpler subproblems and then solve them in sequence. Solving each subproblem is facilitated by the answers to previously solved subproblems.

Our approach differs in two key aspects: (1) our CoT is structurally guided by the path planning algorithm, and (2) our exemplars are automatically selected based on semantic relevance from validation stage. Furthermore, we explore a chat history-based CoT design where previous simplification steps are retained in the conversation context, making the model explicitly aware of its reasoning trajectory.

3 Methodology

3.1 Dynamic Path Planning

Given training, validation, and test sets consisting of single sentences annotated with CEFR levels, we use an LLM to simplify sentences from higher levels to lower levels. Let $L = \{C2, C1, B2, B1, A2, A1\}$ be the set of CEFR levels; $C_{src} \in \{C2, C1, B2\}$ be the source CEFR level; $C_{tgt} \in \{B1, A2, A1\}$ be the target CEFR level. While Barayan et al. (2025) use level-labeled sentences from the training set as prompts to simplify test sentences directly, we apply this approach to validation sentences first. This intermediate step allows us to construct the reward matrix and optimize the simplification path. We use a sentence-level estimator $\phi(\cdot)$ to verify if the generated sentences achieve their target levels. Based on the verification results, we populate the reward matrix R where $R_{i,j}$ represents the effectiveness of simplifying from level i to j:

$$R_{i,j} = \begin{cases} +1, & \text{if achieved level matches target} \\ +0.5, & \text{if differs by one step} \\ -1, & \text{otherwise} \end{cases}$$
(1)

We further normalize the reward scores to the range [0, 1] using min-max normalization. Given this reward matrix, for any source-target level pair, we formulate the path optimization problem as finding the optimal sequence $P^* = [l_0, l_1, \ldots, l_k]$ that maximizes:

$$P^* = argmax_P \sum_{i=1}^{k} R_{l_{i-1}, l_i}$$
(2)

subject to $l_0 = C_{src}$, $l_k = C_{tgt}$, $Level(l_i) < Level(l_{i-1})$.

This optimization is solved using dynamic programming by defining $V(i, t)$ as the maximum reward achievable in t steps ending at level i:

$$V(i, t) = \max_{j < i} \{V(j, t-1) + R_{j,i}\}$$
(3)

The optimal path can then be reconstructed by backtracking through the dynamic programming table.

3.2 Semantic-Guided Exemplar Selection

We begin with the validation set $D_{val} = \{(x_i, l_i)\}_{i=1}^{N}$, where x_i represents a source sentence in D_{val} with its CEFR level l_i in C_{src}, and N the total number of such sentences in D_{val}. We construct exemplar set for each transition step as follows: For each transition (l_s, l_t) in optimal paths (where l_t may be multiple levels below l_s):

- Generate outputs by instructing the LLM to simplify sentences at level l_s to level l_t, with level-labeled sentences from the training set as prompt examples

- Verify the achieved levels using sentence-level estimator $\phi(\cdot)$

- Keep only the successful simplification pairs with exact match:

$$\hat{Y}_{s,t} = \{(x_i, \hat{y}_i) | x_i \in D_{val},$$
$$l_i = l_s, \phi(\hat{y}_i) = l_t\} \quad (4)$$

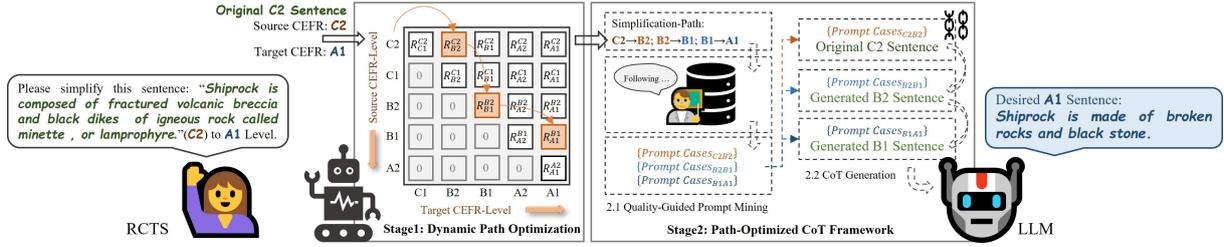where $\hat{y}_i$ is the LLM-generated simplification of x_i.

Figure 1: Research Framework

For each transition (l_s, l_t), we construct the exemplar set $E_{s,t}$ by selecting top k sentence pairs from $\hat{Y}_{s,t}$ with the highest semantic preservation, measured with Semantic Textual Similarity (STS) (Reimers and Gurevych, 2019) between the original sentence x_i and its simplified version $\hat{y}_i$:

$$E_{s,t} = \{(x_i, \hat{y}_i) \in \hat{Y}_{s,t} \mid \text{rank}_{STS}(x_i, \hat{y}_i) \leq k\} \quad (5)$$

In implementation, we set $k = 3$. If $\hat{Y}_{s,t}$ contains fewer than 3 pairs, we use all available pairs as examples. In rare cases when $\hat{Y}_{s,t}$ is empty, we select as examples single sentences from the training set labeled with l_t.

3.3 Few-shot CoT with Chat History

With the optimal path planning and top-k exemplar selection based on semantic preservation, we instruct LLMs to generate simplification step-by-step as illustrated in Figure 1. Barayan et al. (2025) find that optimal performance is achieved through few-shot prompts using a single-case exemplar design that includes target, descriptor, and example. We enhance this approach by incorporating a chat template and chat history that includes the previous simplification steps along the optimal path. Figure 2 shows our prompt design.

Given a test sentence at source proficiency level in C_{src} with target level C_{tgt}, we first identify the optimal path P^* using the reward matrix R. For each transition (l_s, l_t) in P^*, we select exemplar prompt cases from $E_{s,t}$ to guide each step and instruct the LLM to generate a simplification for current transition. We used the generated simplified output $\hat{y}_t$ as the source sentence in the next iteration and append the prompt instruction from previous step to the next prompt design to guide the LLM's CoT generation for the subsequent iteration. After reaching the final step toward the target level C_{tgt}, we use the sentence-level estimator $\phi(\cdot)$ to verify the CEFR level of the final output $\hat{y}_{tgt}$.

This approach uses validated exemplars to en-

sure meaning preservation during simplification. The step-by-step generation process with chat history provides better control over the output quality.

4 Experiment Design

4.1 Dataset

We evaluate our approach on two benchmark datasets. The first is CEFR-SP (Arase et al., 2022), a sentence-level readability assessment corpus for English with CEFR annotations. CEFR-SP comprises three sources: Newsela-Auto (Xu et al., 2015), Wiki-Auto (Jiang et al., 2020), and SCoRe (Chujo et al., 2015). While Barayan et al. (2025) used only the publicly available portions (Wiki-Auto and SCoRe), we obtained additional access to the Newsela-Auto portion through direct contact with Arase et al. (2022). We previously signed agreements with Newsela to use the dataset for research purposes only and will not redistribute it. The second public dataset is README++ (Naous et al., 2024), which contains 9,757 sentences with manual CEFR annotations. These sentences span five languages (Arabic, English, French, Hindi, and Russian) and come from 112 different data sources to ensure domain diversity. We follow the standard train/dev/test splits provided in the official Github repositories of both datasets. The detailed data statistics are provided in Appendix A Table 6.

4.2 Sentence Level Estimator

Our framework requires reliable sentence-level readability assessment for three key functions: reward matrix construction, exemplar selection, and output evaluation. We evaluated two state-of-the-art estimators: the CEFR-SP-estimator provided with the CEFR-SP dataset for English texts, and the README-estimator developed by Naous et al., 2024 which supports all five languages in their dataset.

To ensure optimal selection of sentence-level readability estimator, we conducted comprehensive

Figure 2: Prompt Design for Few-shot CoT with Chat Template and Chat History

benchmarking against various sentence readability assessment baselines, including traditional readability metrics and RSRS (Martinc et al., 2021) implementations with different language models. Our evaluations on both CEFR-SP and README++ datasets (Appendix C (Tables 8 and 9)) led us to adopt the CEFR-SP-estimator for CEFR-SP experiments and the README-estimator for README++ experiments, as they demonstrated superior performance on their respective datasets.

4.3 Instruction-Tuned LLM

We evaluated three open-source LLMs: Llama-3-8B-Instruct (Dubey et al., 2024), Gemma-2-9b-it (Team et al., 2024), and Mistral-7B-Instruct-v0.3 (Jiang et al., 2023). We find a consistent tradeoff pattern between simplification effectiveness and meaning preservation across all three LLMs. However, Mistral's performance is quite erratic. Therefore, we focus our results analysis on Llama and Gemma.

4.4 Automatic Evaluation

We use the following automatic metrics for simplification effectiveness and meaning preservation.
Spearman Correlation (ρ): Rank correlation between target and estimated CEFR levels.
RMSE: Root mean squared error between estimated and target CEFR levels.
Adjacent Accuracy: Percentage of outputs within one CEFR level of the target.
Exact Accuracy: Percentage of output matching exactly the level of the target.
Semantic Textual Similarity (STS): Cosine similarity between original and simplified texts, computed using sentence transformers (Reimers and

Gurevych, 2019). We use all-MiniLM-L6-v2 for English and paraphrase-multilingual-MiniLM-L12-v2 for the other languages.
BertScore: Token-level semantic similarity between original and simplified texts, computed using BERT contextual embeddings (Zhang et al., 2020).

4.5 Human Evaluation

We further conduct comprehensive human evaluations on readability/proficiency and meaning preservation.
Readability Assessment For each language, we recruited three evaluators, totaling 15 evaluators across five languages. All evaluators were non-native speakers with extensive university-level foreign language teaching experience. The English evaluators averaged 26 years of teaching English as a second language, French evaluators 13 years, Russian evaluators 6 years, Arabic evaluators 11 years, and Hindi evaluators 8 years. To ensure familiarity with the CEFR descriptor scheme, we implemented a qualification task. For each language, we selected 30 sentences from benchmark datasets previously annotated with CEFR levels (10 each from levels B1, A2, and A1). We evaluated the annotators' CEFR assessments against gold references using three metrics: Majority Agreement, Krippendorff's Alpha, and Gwet AC2.

Meaning Preservation Judgement For each language, the three instructor evaluators annotated 45 sentence pairs (original vs. simplified) across 9 source-target transitions for meaning preservation. Following Barayan et al. (2025), we employed a scale where 0 indicates no or trivial changes; 1 signifies non-trivial changes with the main idea maintained; and 2 indicates failure to preserve the

main idea. Intermediate scores (0.5 and 1.5) were also permitted.

From the 45 annotated pairs, we selected 20 pairs showing high agreement among instructor evaluators. High agreement was determined when at least two of our three metrics (Majority Agreement, Krippendorff's Alpha, and Gwet AC2) exceeded fair value thresholds. For these pairs, we used the majority judgment of the three annotators to establish meaning preservation scores, creating a qualification test for subsequent assessments.

We then recruited three native speakers for each language, all with college-level or higher education, to evaluate meaning preservation using these 20 sentence pairs. However, for French and Hindi, neither consistent agreement among the native speakers nor alignment with instructor reference scores was achieved. Consequently, for these two languages, we relied on the instructors for meaning preservation annotation. For English and Russian, only one native speaker per language achieved high agreement with instructor reference scores, so we paired these native speakers with two instructors to complete the annotation task. Similarly, for Arabic, we paired two native speakers with one instructor due to agreement. We pay $16 to $18 per hour for human annotation.

4.6 Baselines

We compare our research framework with the following baselines:

COPY: A simple baseline that returns the input sentence unchanged, serving as a lower bound for performance evaluation.

Supervised: An mT5-base model fine-tuned on available parallel simplification data for each language (details in Appendix D.2). For *Supv. mT5-base (lang)*, mT5-base was fine-tuned on language-specific parallel data in the README dataset. For *Supv. mT5-base (all)*, mT5-base was fine-tuned on combined parallel data available in the README dataset. Since CEFR dataset is English-only, we denote the supervised model as *Supv. mT5-base* without further specification.

Barayan et al. (2025) : State-of-the-art results on the publicly available CEFR-SP dataset (Wiki-Auto and SCoRE portions), as reported in their paper.

One-step single case: Our reimplementation of Barayan et al. (2025) with one-step LLM generation and single-case prompt without semantic exemplars. We use their few-shot prompting without chat template. We run three sets of experiments

with temperature of 1 and report average performances. Referencing the result Table 1 and 7, this model is configured as *Llama_w/o_chat+N+N+N*.
One-step single Chat: We implement Barayan et al. (2025) with one-step LLM generation and prompt without semantic exemplars. But we use chat template without chat history. In the result Tables 1 and 7, this model is configured as *Model+N+N+N*.

For our proposed framework, we experiment with the following models referencing notation in result Tables 1 and 7.

One-step Semantic Chat which is One-step LLM generation with semantic-preservation exemplars selection and chat template. The configuration is denoted as *Model+N+Y+N*.

Planning+Single-case CoT Chat History: Two-stage approach using path planning and CoT generation with single-case exemplars and chat template and history. The configuration is denoted as *Model+Y+N+Y*.

Planning+Semantic CoT No History: Two-stage approach using path planning and CoT generation with semantic exemplars and chat template without history. The configuration is denoted as *Model+Y+Y+N*.

Planning+Semantic Chat History: Our complete framework combining path planning with semantic-aware CoT and chat history. The configuration is denoted as *Model+Y+Y+Y*.

5 Results and Analysis

Due to space constraints, we present results on README-EN in the main text (Table 1) and provide comprehensive results across all datasets in Appendix Table 7.

5.1 Trade-off Between Simplification Effectiveness and Meaning Preservation

Comparing Llama and Gemma models across all experiments, a clear trade-off pattern emerges. On README-EN, Gemma's Planning+Single case+CoT configuration achieves the highest adjacent accuracy (96.46%) but lower STS (61.32%), while Llama's same configuration reaches 93.43% adjacent accuracy with higher STS (65.81%). This inverse relationship between target level achievement and semantic preservation is consistent across both datasets. Llama is generally better at semantic preservation than Gemma but generally underperforms relative to Gemma in simplification.

Model	Planning	Semantic	History	ρ ↑	AdjAcc (%) ↑	ExactAcc (%) ↑	RMSE ↓	BS_F1 ↑	STS (%) ↑
COPY	–	–	–	0.00	32.32	6.82	2.29	100.0	100.0
Supv. mT5-base (lang)	–	–	–	0.01	45.96	14.39	1.99	94.00	84.05
Supv. mT5-base (all)	–	–	–	0.03	42.68	10.86	2.03	96.88	92.36
Llama_w/o_chat	N	N	N	0.12	62.62	24.66	1.60	91.05	70.41
Llama	N	N	N	0.45	72.98	26.26	1.27	90.72	73.43
Llama	N	Y	N	0.49	76.52	23.23	1.26	91.44	76.02
Llama	Y	N	Y	0.68	93.43	46.46	0.86	89.15	65.81
Llama	Y	Y	N	0.37	70.96	21.97	1.37	91.60	78.01
Llama	Y	Y	Y	0.66	91.41	37.37	0.95	90.00	69.90
Gemma	N	N	N	0.43	83.33	31.06	1.13	89.79	70.11
Gemma	N	Y	N	0.48	86.62	30.05	1.08	89.93	69.67
Gemma	Y	N	Y	0.66	96.46	44.70	0.81	88.10	61.32
Gemma	Y	Y	N	0.45	85.35	33.39	1.08	89.54	68.66
Gemma	Y	Y	Y	0.58	91.92	37.63	0.94	88.97	66.49

Table 1: Comprehensive results for README-EN.

5.2 Dynamic Path Planning Improves Readability Effectiveness

Introducing dynamic path planning substantially boosts simplification effectiveness for both models. Comparing "Plan+single CoT" with "One-step single chat" configurations shows the impact of multi-step simplification. On README-EN, Llama's exact accuracy improves from 26.26% (one-step single chat) to 46.46% (Plan+single CoT with history), while Spearman correlation increases from 0.45 to 0.68. Similarly on CEFR-SP-Partial, Gemma improves from $\rho = 0.31$ (one-step) to 0.59 (planning with history). This effectiveness gain comes with a trade-off in semantic preservation. For instance, on CEFR-SP-Partial with Llama, STS decreases from 79.56% (one-step) to 71.46% (planning), reflecting the inherent tension between aggressive simplification and meaning preservation. The inherent difficulty of large-span simplification is illustrated by representative examples in Appendix H.

5.3 Chat History and Semantic Exemplars Optimize the Effectiveness-Preservation Balance

Adding chat history which retains previous simplification steps further enhances level accuracy and stability. By making the LLM explicitly aware of its reasoning trajectory, chat history enables more coherent multi-step simplification. Across datasets, including history raises ρ and adjacent accuracy by 3-6 points relative to the no-history variant. For example, on README-EN with Llama, "Plan+Semantic CoT with history" achieves $\rho = 0.66$ and adjacent accuracy of 91.41%, compared to $\rho = 0.37$ and 70.96% without history.

Semantic-aware exemplar selection provides complementary benefits for meaning preservation. Comparing "Plan+Semantic CoT" versus "Plan+single CoT" configurations reveals consistent improvements in semantic fidelity. On CEFR-SP-Partial with Llama, incorporating semantic exemplars improves STS from 71.46% (single case) to 75.35% (semantic selection), while maintaining strong readability control ($\rho = 0.56$). This demonstrates that selecting exemplars based on meaning preservation helps guide the LLM toward simplifications that better retain semantic content.

Among all tested configurations, Planning + Semantic CoT + Chat History achieves the best overall balance—offering the highest readability control for Gemma and the strongest meaning preservation for Llama—demonstrating that dynamic, context-aware generation provides a robust framework for readability-controlled simplification.

5.4 Human Evaluation Results

Inter-annotator reliability. Table 2 shows acceptable inter-annotator agreement across all languages, with Krippendorff's Alpha ranging from 0.62-0.79 and Gwet AC2 from 0.5-0.68.

Readability Control. Table 3 demonstrates strong simplification effectiveness of our Planning+Semantic CoT approach, achieving high Spearman correlation ($\rho = 0.49$-0.66) and adjacent accuracy (95.45-100%) across all languages.

Human Meaning Preservation Judgement Table 4 reveals a critical finding: the trade-off between simplification effectiveness and meaning preservation observed in automatic evaluation (Section 5.1) is confirmed by human assessment. While one-step simplification maintains better semantic preservation (average scores 0.71-1.14), our Planning+Semantic CoT method shows higher meaning loss (0.93-1.47), particularly for large-span simplifications (C2 to A1). Notably, establishing reliable meaning preservation judgments proved exceptionally challenging. Despite following established protocols (Section 4.5), we encountered unexpected difficulties in achieving annotator agreement. Surprisingly, native speakers showed even

Language	Majority Agreement	Krippendorff's Alpha	Gwt AC2
English	91.11%	0.62	0.5
French	93.33%	0.63	0.55
Russian	96.67%	0.73	0.67
Arabic	100%	0.7	0.59
Hindi	100%	0.79	0.68

Table 2: Human Readability Assessment: Inter-annotator Agreement

Language	ρ	Adj Acc (%)	Exa Acc (%)	RMSE
English	0.66	97.78	62.22	0.67
French	0.64	97.78	60	0.68
Russian	0.65	100	53.33	0.68
Arabic	0.63	97.78	60	0.68
Hindi	0.49	95.45	54.55	0.77

Table 3: Human Readability Assessment for Planning+Semantic Chat History

greater disagreement among themselves than language instructors. For French and Hindi, we ultimately relied on instructor judgments due to insufficient native speaker consensus. This annotation challenge underscores a fundamental difficulty: if human experts struggle to agree on meaning preservation, it highlights the inherent complexity LLMs face when attempting to maintain semantic fidelity during large-span simplifications.

5.5 Efficiency of Path Planning

To further evaluate the efficiency of dynamic path planning, we also perform an experiment where the simplification steps are sequential from source to target, with semantic CoT and chat history, to compare with Plan+Semantic CoT with Chat History. Evaluation shows that the simplification effectiveness scores and the meaning preservation scores are very similar between these two, indicating almost identical performances. Therefore, we examine the efficiency of path planning by measuring the "reduction ratio of inference steps" of Plan+Semantic CoT with Chat History compared to Sequential+Semantic CoT with Chat History. We present Llama results in Table 5 and Gemma results in Appendix Table 11. Results show that the dynamic path planning approach achieves a substantial reduction in computational overhead, decreasing the required processing steps by 22-42%.

6 Conclusions

We presented a novel framework for multilingual proficiency-controlled sentence simplification that addresses the fundamental challenge of simplifying across large readability spans. Our approach combines dynamic path planning, semantic-aware ex-

emplar selection to preserve meaning, and chain-of-thought generation with chat history for coherent multi-step reasoning. Comprehensive evaluation across five languages demonstrates that our framework significantly improves target level achievement compared to single-step approaches. Dynamic path planning also reduces computational steps by 22-42% through efficient path optimization. However, our results reveal an inherent trade-off: achieving precise readability control, especially for large-span simplifications, comes at the cost of semantic preservation. This challenge is validated by both automatic metrics and human evaluation, where even human annotators struggle to agree on meaning preservation judgments—highlighting the fundamental difficulty of this task. (see Appendix H for illustrative examples of the complexity gap between C2 and A1 levels). Future work should explore techniques to better balance this trade-off.

7 Limitations

While our approach shows promising results, several important limitations require attention. First, we focus on sentence-level simplification, while real-world applications also require document-level coherence and cross-sentence discourse relationships. Second, our human evaluation revealed fundamental challenges in achieving consensus on meaning preservation, suggesting that current evaluation paradigms may inadequately capture semantic fidelity. Third, the trade-off between readability control and meaning preservation intensifies with larger simplification spans, indicating that optimizing both dimensions simultaneously remains an open challenge. The writing of this paper was partially assisted by Claude for editing and refinement.

8 Acknowledgments

Language	Model	Avg. Score (Stdv)	C2 simplified to:			C1 simplified to:			B2 simplified to:		
			B1	A2	A1	B1	A2	A1	B1	A2	A1
English	Llama One-step single chat	0.822 (0.327)	0.6	0.6	1	0.6	0.4	0.9	0.7	1.2	1.4
	Llama Plan+Semantic CoT (with hist.)	1.089 (0.71)	0.1	1.2	1.9	0.5	1.5	2	0.3	0.7	1.6
French	Llama One-step single chat	0.889 (0.513)	0.5	1.5	1.3	0.4	0.4	0.6	0.9	1.5	1.2
	Llama Plan+Semantic CoT (with hist.)	0.933 (0.308)	0.3	0.7	0.9	1	1.2	1.3	1	0.8	1.2
Russian	Llama One-step single chat	0.711 (0.257)	0.5	0.8	1	0.6	0.6	0.9	0.6	1.1	0.3
	Llama Plan+Semantic CoT (with hist.)	1.133 (0.492)	0.9	1.3	1.8	0.4	1.1	1.7	0.5	1.0	1.5
Arabic	Llama One-step single chat	1.144 (0.378)	0.9	0.7	1	1.1	0.9	1.2	1.6	1.9	1.0
	Llama Plan+Semantic CoT (with hist.)	1.422 (0.441)	0.6	1.3	2	0.9	1.4	1.8	1.5	1.6	1.7
Hindi	Llama One-step single chat	1.089 (0.386)	1.0	0.9	1	1.3	1.1	1.3	1.9	0.6	0.7
	Llama Plan+Semantic CoT (with hist.)	1.467 (0.52)	2.0	1.2	1.9	1.6	0.9	2.0	0.7	1.0	1.9

Table 4: Human Judgement of Meaning Preservation

Corpus	Source	Target	DP-Path	RRIS↑
CEFR-SP-Partial	C2	B1	[B2, B1]	25.0%
	C2	A2	[B2, B1, A2]	
	C2	A1	[B2, B1, A2, A1]	
CEFR-SP-Whole	C2	B1	[B2, B1]	25.0%
	C2	A2	[B2, B1, A2]	
	C2	A1	[B2, B1, A2, A1]	
README_EN	C2	B1	[B2, B1]	25.0%
	C2	A2	[B2, B1, A2]	
	C2	A1	[B2, B1, A2, A1]	
README_FR	C2	A2	[B2, B1, A2]	22.2%
	C2	A1	[B2, B1, A2, A1]	
README_RU	C2	B1	[B2, B1]	25.0%
	C2	A2	[B2, B1, A2]	
	C2	A1	[B2, B1, A2, A1]	

Table 5: RRIS on Llama. RRIS denotes the **R**eduction **R**atio of **I**nference **S**teps relative to the Sequential-steps strategy.

References

Sweta Agrawal and Marine Carpuat. 2023. Controlling pre-trained language models for grade-specific text simplification. In *Proceedings of ACL 2023*.

Fernando Alva-Manchego, Joachim Bingel, Gustavo Paetzold, Carolina Scarton, and Lucia Specia. 2017. Learning how to simplify from explicit labeling of complex-simplified text pairs. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 295–305, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Yuki Arase, Satoru Uchida, and Tomoyuki Kajiwara. 2022. Cefr-based sentence difficulty annotation and assessment. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6206–6219.

Abdulelah Barayan, Jose Camacho-Collados, and Fernando Alva-Manchego. 2025. Analysing zero-shot readability-controlled sentence simplification. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6762–6781.

Sofia Blinova, Xinyu Zhou, Martin Jaggi, Carsten Eickhoff, and Seyed Ali Bahrainian. 2023. SIMSUM: Document-level text simplification via simultaneous summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9927–9944, Toronto, Canada. Association for Computational Linguistics.

Rémi Cardon and Natalia Grabar. 2020. French biomedical text simplification: When small and precise helps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 710–716, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Kiyomi Chujo, Kathryn Oghigian, and Shiro Akasegawa. 2015. A corpus and grammatical browsing system for remedial efl learners. In Agnieszka Lenko-Szymańska and Alex Boulton, editors, *Multiple Affordances of Language Corpora for Data-driven Learning*, volume 69 of *Studies in Corpus Linguistics*, pages 109–128. John Benjamins.

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

Yue Dong, Zichao Li, Mehdi Rezagholizadeh, and Jackie Chi Kit Cheung. 2019. Editnts: A neural programmer-interpreter model for sentence simplification through explicit editing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3393–3402, Florence, Italy. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Arian Farajidizaji, Vinit Raina, and Mark Gales. 2024. Is it possible to modify text to a target readability level? an initial investigation using zero-shot large language models. *arXiv preprint arXiv:2309.12551*.

Goran Glavaš and Sanja Štajner. 2015. Simplifying lexical simplification: Do we need simplified corpora? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 63–68, Beijing, China. Association for Computational Linguistics.

Natalia Grabar and Rémi Cardon. 2018. CLEAR – simple corpus for medical French. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, pages 3–9, Tilburg, the Netherlands. Association for Computational Linguistics.

Joseph Marvin Imperial, Abdullah Barayan, Regina Stodden, Rodrigo Wilkens, Ricardo Munoz Sanchez, Lingyun Gao, Melissa Torgbi, Dawn Knight, Gail Forey, Reka R. Jablonkai, Ekaterina Kochmar, Robert Reynolds, Eugénio Ribeiro, Horacio Saggion, Elena Volodina, Sowmya Vajjala, Thomas François, Fernando Alva-Manchego, and Harish Tayyar Madabushi. 2025. Universalcefr: Enabling open multilingual research on language proficiency assessment. *Preprint*, arXiv:2506.01419.

Joseph Marvin Imperial and Harish Tayyar Madabushi. 2023. Flesch or fumble? evaluating readability standard alignment of instruction-tuned language models. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 205–223.

L Javourey-Drevet, S Dufau, T François, N Gala, J Ginestié, and JC Ziegler. 2022. Simplification of literary and scientific texts to improve reading fluency and comprehension in beginning readers of French. *Applied Psycholinguistics*, 43(2):485–512.

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

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.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 22199–22213. Curran Associates, Inc.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, William B. Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for gpt-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland. Association for Computational Linguistics.

Mounica Maddela, Fernando Alva-Manchego, and Wei Xu. 2021. Controllable text simplification with explicit paraphrasing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 3536–3553, Online. Association for Computational Linguistics.

Ali Malik, Stephen Mayhew, Chris Piech, and Klinton Bicknell. 2024. From tarzan to tolkien: Controlling the language proficiency level of llms for content generation. *Preprint*, arXiv:2406.03030.

Louis Martin, Éric Villemonte de la Clergerie, Benoît Sagot, and Antoine Bordes. 2020. Controllable sentence simplification. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC)*, pages 4689–4698, 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.

Kaijie Mo and Renfen Hu. 2024. ExpertEase: A multi-agent framework for grade-specific document simplification with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9080–9099, Miami, Florida, USA. Association for Computational Linguistics.

Amit Moryossef, Yoav Goldberg, and Ido Dagan. 2019. Step-by-step: Separating planning from realization in neural data-to-text generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2267–2277, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Tarek Naous, Michael J Ryan, Anton Lavrouk, Mukesh Chandra, and Wei Xu. 2024. Readme++: Benchmarking multilingual language models for multidomain readability assessment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12230–12266.

Masashi Oshika, Makoto Morishita, Tsutomu Hirao, Ryohei Sasano, and Koichi Takeda. 2024. Simplifying translations for children: Iterative simplification considering age of acquisition with llms. *Preprint*, arXiv:2408.04217.

Gustavo Paetzold and Lucia Specia. 2015. Lexenstein: A framework for lexical simplification. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 85–90, Beijing, China. Association for Computational Linguistics.

Ratish Puduppully, Li Dong, and Mirella Lapata. 2019. Data-to-text generation with content selection and planning. In *Proceedings of the AAAI Conference on*

Artificial Intelligence, volume 33, pages 6908–6915, Honolulu, Hawaii, USA. AAAI Press.

Jipeng Qiang, Yun Li, Yi Zhu, Yunhao Yuan, Yang Shi, and Xindong Wu. 2021. Lsbert: Lexical simplification based on bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3064–3076.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

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

Andrey Sakhovskiy, Alexandra Izhevskaya, Alena Pestova, Elena Tutubalina, Valentin Malykh, Ivana Smurov, and Ekaterina Artemova. 2021. RuSimpleSentEval-2021 shared task: Evaluating sentence simplification for russian. In *Proceedings of the International Conference "Dialogue"*, pages 607–617.

Carolina Scarton and Lucia Specia. 2018. Learning simplifications for specific target audiences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 712–718.

Kevin Stowe, Debanjan Ghosh, and Mengxuan Zhao. 2022. Controlled language generation for language learning items. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track (EMNLP Industry Track)*, pages 294–305, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu,

Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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.

Bohan Zhang, Prafulla Kumar Choubey, and Ruihong Huang. 2022. Predicting sentence deletions for text simplification using a functional discourse structure. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 255–261, Dublin, Ireland. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yang Zhong, Chao Jiang, Wei Xu, and Junyi Jessy Li. 2020. Discourse level factors for sentence deletion in text simplification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9709–9716, New York, New York, USA. AAAI Press.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, et al. 2023. Least-to-most prompting enables complex reasoning in large language models. In *Proceedings of the Eleventh International Conference on Learning Representations (ICLR)*. OpenReview.net.

A Data Statistics

We follow the standard train/dev/test splits from the benchmark datasets CEFR-SP and README++, as detailed in Table 6. For both datasets, since the test sets are used to evaluate simplification from higher to lower levels, they only contain sentences from higher CEFR levels (B2, C1, and C2).

B Comprehensive Results Across All Datasets

Table 7 contains full experimental results for all language-dataset combinations, including details for each model configuration.

	Split	A1	A2	B1	B2	C1	C2	Total
CEFR-SP (Partial)	Train	47	959	2245	2472	1364	91	7178
	Dev	40	154	513	399	193	67	1366
	Test	-	-	-	142	187	72	401
CEFR-SP (Whole)	Train	55	1324	5481	5209	1821	100	13990
	Dev	41	165	685	651	227	74	1843
	Test	-	-	-	188	227	74	489
README_EN	Train	146	539	485	713	304	56	2243
	Dev	14	69	61	91	39	9	283
	Test	-	-	-	92	34	6	132
README_FR	Train	107	302	343	277	193	97	1319
	Dev	14	31	52	29	22	17	165
	Test	-	-	-	46	26	12	84
README_RU	Train	321	223	334	260	193	73	1404
	Dev	39	34	38	32	23	10	176
	Test	-	-	-	34	21	8	63
README_AR	Train	67	198	414	434	284	146	1543
	Dev	6	26	44	63	38	19	196
	Test	-	-	-	68	28	18	114
README_HI	Train	204	230	223	200	172	150	1179
	Dev	27	22	38	31	20	11	149
	Test	-	-	-	32	30	13	75

Table 6: Data Statistics

C Comparing Sentence Level Estimators

C.1 CEFR-SP

We conducted extensive comparisons of readability assessment methods across both datasets. Table 8 presents the evaluation of different approaches on CEFR-SP data, comparing unsupervised models (RSRS trained with BERT and XLM-R variants), traditional metrics (FKGL, ARI), and CEFR-SP-estimator and README-estimator. Traditional metrics showed limited correlation (FKGL: 0.734, ARI: 0.71), while neural RSRS (Martinc et al., 2021) implementations achieved moderate performance (RSRS_XLMR_B: 0.678). The CEFR-SP estimator demonstrated the strongest performance (0.937).

C.2 README

Table 9 compares README estimator and CEFR estimator on the README dataset. Although README-estimator achieves 0.00 F1 on category C2, most of its predictions fall into the adjacent category C1, resulting in minimal practical error. The consistent superior F1 over all levels except for C2 and the higher Average F1 score (49.59 vs. 47.9) justify selecting README-estimator over CEFR-SP-estimator.

D Model Parameters

D.1 LLM Inference Setting

We conduct the inference process on an NVIDIA GeForce RTX 3080, leveraging the Transformers library (Wolf et al., 2020). For generation, we set a maximum output length of 128 tokens, a temperature of 0.0 and do_sample=False to ensure reproducibility.

D.2 Supervised mt5 Training and Inference Setting

For the supervised baseline model training, the batch size was 4, and the optimizer used was Adam (Kingma and Ba, 2017) with learning rate of 2e-5. For $Supv.mT5 - base(lang)$, mT5-base was fine-tuned on language-specific parallel data only in the README dataset. Since CEFR dataset is English-only, we denote the supervised model as $Supv.mT5 - base$ without further specification. We configure the validation step to be 500 and conducted the validation process 10 times. As for $Supv.mT5 - base(all)$, mT5-base was fine-tuned on combined parallel data available in the README dataset. the validation step was set to 1500, also with 10 repetitions of the validation. Subsequently, we adopted SARI (Xu et al., 2016) and selected the checkpoint achieving the highest SARI score on the validation set as the final model.

The total GPU computation time was under 4 hours for mt5 training and inference.

E Supervised Baseline Training and Inference

For the supervised baseline experiments, we constructed parallel datasets for different languages following Ryan et al. (2023):

English: We followed the methodology from Barayan et al. (2025), though our train/dev split distribution differs slightly as the original split details were not available.

French: We combined data from two sources: Wik-iLarge (Cardon and Grabar, 2020, randomly selected 5,000 sentence pairs) and CLEAR (Grabar and Cardon, 2018, included all available sentence pairs). The distribution of this data across CEFR levels is shown in Table 5.

Russian: Utilized the complete RSSE (Sakhovskiy et al., 2021) dataset.

For Hindi and Arabic, no publicly available parallel sentence datasets were available for text simplification.

We implemented two training strategies. The first approach trained individual models for each language and generated simplifications on the validation set. The second strategy fine-tuned mT5-base using combined English, French, and Russian data, then generated simplifications on validation sets. For implementation, we used mT5-base as our base model. All sentences were prepended with their corresponding CEFR level during training and validation. The parallel sentence pairs consisted of source and simplified versions, with validation performed on language-specific validation sets. Table 10 presents the detailed distribution of training and validation data across CEFR levels.

F Human Evaluation Instruction Screenshots

Figures 3 and 4 provide screenshots of the user interface and detailed instructions given to human annotators for both readability assessment and meaning preservation tasks.

G Efficiency of Path Planning

Table 11 compares dynamic path planning against sequential simplification based on Gemma, including inference step reduction ratios across different languages.

Figure 3: Screenshot of Human Meaning Preservation Judgement: Instruction

Figure 4: Screenshot of Human Meaning Preservation Judgement: Annotation

H Illustrative Examples of CEFR Level Complexity

To illustrate the inherent difficulty of large-span proficiency-controlled simplification, we present representative examples from our benchmark datasets showing the stark contrast between beginner (A1) and expert (C2) level texts.

A1 Level Examples:

- "How was your first day in college?"

- "That's a good question."

- "And I can tell you this, this can be done."

A1 represents the beginner level where sentences are short, use common everyday vocabulary, simple grammatical structures, and express concrete, straightforward ideas.

C2 Level Examples:

Technical domain:

"This is referred to as the dual space-time variation property and it indicates that precoding using arbitrary joint space-time orthogonal basis is not sufficient to ensure interference-free communication, unless these basis remain orthogonal after propagating through the channel."

Literary domain:

"Would you allow me, in my gratitude for the benevolent reception that you gave me one day, to draw the attention of your rightful glory and to tell you that your star, so happy until now, is threatened by the most shameful and most ineffaceable of blemishes?"

C2 represents mastery level where sentences contain domain-specific terminology, complex syntactic structures, abstract concepts, and sophisticated lexical choices.

The Inherent Challenge: These examples illustrate why simplification from C2 to A1 entails inevitable trade-offs. Transforming such content to beginner-level comprehensibility requires compression, abstraction, and semantic loss. Our framework does not eliminate this fundamental trade-off but rather makes it controllable and optimized through multi-step planning and semantic-aware generation.

Dataset	Model	Planning	Semantic	History	ρ ↑	AdjAcc (%) ↑	ExactAcc (%) ↑	RMSE ↓	BS_F1 ↑	STS (%) ↑
CEFR-SP-Part	COPY	–	–	–	0.00	24.15	3.99	2.45	100.0	100.0
	Supv. mT5-base	–	–	–	0.01	39.43	–	2.08	65.15	84.51
	Barayan et al., 2025	N	N	N	0.19	48.21	–	1.80	55.28	82.86
	Llama_w/o_chat	N	N	N	0.24	54.2	16.87	1.77	41.94	73.03
	Llama	N	N	N	0.28	53.28	16.46	1.74	51.39	79.56
	Llama	N	Y	N	0.35	48.13	12.55	1.81	57.69	83.30
	Llama	Y	N	Y	0.52	74.73	30.01	1.35	39.66	71.46
	Llama	Y	Y	N	0.41	51.29	12.55	1.80	55.51	82.71
	Llama	Y	Y	Y	0.56	68.99	26.43	1.49	45.94	75.35
	Gemma	N	N	N	0.31	62.34	21.86	1.60	44.69	75.65
	Gemma	N	Y	N	0.40	63.67	20.45	1.56	46.60	76.55
	Gemma	Y	N	Y	0.59	83.87	38.99	1.13	32.36	65.66
	Gemma	Y	Y	N	0.36	62.34	20.70	1.59	44.35	75.73
	Gemma	Y	Y	Y	0.49	72.07	28.10	1.42	39.28	72.77
CEFR-SP-Whole	COPY	–	–	–	0.00	26.52	4.77	2.51	100.0	100.0
	Supv. mT5-base	–	–	–	0.00	38.85	11.52	2.20	89.17	70.57
	Llama_w/o_chat	N	N	N	0.06	47.31	16.54	2.05	58.77	78.03
	Llama	N	N	N	0.29	55.08	17.45	1.69	51.90	79.49
	Llama	N	Y	N	0.35	50.72	13.84	1.76	57.68	82.62
	Llama	Y	N	Y	0.53	77.51	31.42	1.29	40.04	71.11
	Llama	Y	Y	N	0.37	52.01	13.77	1.78	56.57	83.06
	Llama	Y	Y	Y	0.55	70.69	27.33	1.46	47.12	75.68
	Gemma	N	N	N	0.32	64.62	23.31	1.54	45.19	75.25
	Gemma	N	Y	N	0.36	67.01	24.40	1.51	45.29	75.14
	Gemma	Y	N	Y	0.19	50.65	16.56	1.85	32.70	65.53
	Gemma	Y	Y	N	0.36	66.39	24.06	1.49	41.97	72.40
	Gemma	Y	Y	Y	0.51	78.05	32.31	1.28	37.06	69.34
README-AR	COPY	–	–	–	0.00	26.02	4.68	2.61	100.0	100.0
	Supv. mT5-base (all)	–	–	–	0.01	55.56	21.35	1.95	90.63	75.39
	Llama_w/o_chat	N	N	N	0.11	44.93	9.26	1.91	85.56	66.37
	Llama	N	N	N	0.18	48.54	15.50	1.94	91.79	81.61
	Llama	N	Y	N	0.24	48.25	15.20	1.95	92.59	83.81
	Llama	Y	N	Y	0.53	73.39	20.47	1.39	89.56	71.27
	Llama	Y	Y	N	0.27	54.09	17.54	1.76	91.86	80.22
	Llama	Y	Y	Y	0.58	74.27	17.54	1.39	90.70	74.38
	Gemma	N	N	N	0.25	55.85	18.71	1.73	89.90	74.89
	Gemma	N	Y	N	0.26	59.06	18.42	1.69	90.08	75.61
	Gemma	Y	N	Y	0.51	78.65	24.56	1.33	88.30	66.96
	Gemma	Y	Y	N	0.28	62.57	21.05	1.59	89.77	74.49
	Gemma	Y	Y	Y	0.45	69.30	20.76	1.50	89.38	72.27
README-EN	COPY	–	–	–	0.00	32.32	6.82	2.29	100.0	100.0
	Supv. mT5-base (lang)	–	–	–	0.01	45.96	14.39	1.99	94.00	84.05
	Supv. mT5-base (all)	–	–	–	0.03	42.68	10.86	2.03	96.88	92.36
	Llama_w/o_chat	N	N	N	0.12	62.62	24.66	1.60	91.05	70.41
	Llama	N	N	N	0.45	72.98	26.26	1.27	90.72	73.43
	Llama	N	Y	N	0.49	76.52	23.23	1.26	91.44	76.02
	Llama	Y	N	Y	0.68	93.43	46.46	0.86	89.15	65.81
	Llama	Y	Y	N	0.37	70.96	21.97	1.37	91.60	78.01
	Llama	Y	Y	Y	0.66	91.41	37.37	0.95	90.00	69.90
	Gemma	N	N	N	0.43	83.33	31.06	1.13	89.79	70.11
	Gemma	N	Y	N	0.48	86.62	30.05	1.08	89.93	69.67
	Gemma	Y	N	Y	0.66	96.46	44.70	0.81	88.10	61.32
	Gemma	Y	Y	N	0.45	85.35	33.39	1.08	89.54	68.66
	Gemma	Y	Y	Y	0.58	91.92	37.63	0.94	88.97	66.49
README-FR	COPY	–	–	–	0.00	28.57	6.75	2.42	100.0	100.0
	Supv. mT5-base (lang)	–	–	–	0.12	75.79	21.83	1.39	84.16	43.76
	Supv. mT5-base (all)	–	–	–	0.05	44.44	14.29	1.99	95.56	88.72
	Llama_w/o_chat	N	N	N	0.22	57.28	19.44	1.66	89.55	76.93
	Llama	N	N	N	0.28	48.81	14.29	1.79	93.22	85.77
	Llama	N	Y	N	0.31	51.59	12.70	1.76	93.45	86.08
	Llama	Y	N	Y	0.67	79.76	20.63	1.21	91.03	76.38
	Llama	Y	Y	N	0.38	51.59	13.10	1.70	92.91	84.77
	Llama	Y	Y	Y	0.68	76.98	17.46	1.26	91.75	78.23
	Gemma	N	N	N	0.53	63.49	18.65	1.43	91.71	80.74
	Gemma	N	Y	N	0.50	66.27	21.03	1.38	91.77	80.19
	Gemma	Y	N	Y	0.75	89.68	27.78	1.03	89.69	69.35
	Gemma	Y	Y	N	0.52	69.84	21.83	1.36	91.14	77.14
	Gemma	Y	Y	Y	0.66	81.75	25.79	1.14	90.65	74.93
README-RU	COPY	–	–	–	0.00	18.52	4.23	2.73	100.0	100.0
	Supv. mT5-base (lang)	–	–	–	0.00	64.02	23.28	1.60	92.33	79.77
	Supv. mT5-base (all)	–	–	–	0.08	52.38	17.46	1.91	94.18	87.25
	Llama_w/o_chat	N	N	N	0.29	55.38	19.75	1.83	89.62	75.90
	Llama	N	N	N	0.34	68.78	23.81	1.45	92.60	84.73
	Llama	N	Y	N	0.43	68.25	17.99	1.48	93.09	86.20
	Llama	Y	N	Y	0.76	97.35	59.26	0.70	90.36	73.59
	Llama	Y	Y	N	0.41	67.72	20.63	1.42	92.47	83.65
	Llama	Y	Y	Y	0.73	93.12	49.74	0.84	91.15	76.19
	Gemma	N	N	N	0.53	85.71	41.80	1.03	91.12	78.35
	Gemma	N	Y	N	0.51	84.13	32.28	1.16	91.75	81.14
	Gemma	Y	N	Y	0.77	99.47	64.02	0.61	89.65	69.01
	Gemma	Y	Y	N	0.64	92.06	45.50	0.90	91.18	78.35
	Gemma	Y	Y	Y	0.70	93.12	50.26	0.84	90.72	75.06
README-HI	COPY	–	–	–	0.00	28.00	8.00	3.09	100.0	100.0
	Supv. mT5-base (all)	–	–	–	0.01	51.11	14.67	2.24	93.04	75.19
	Llama_w/o_chat	N	N	N	0.13	48.74	18.66	2.30	89.98	74.63
	Llama	N	N	N	0.21	45.33	12.44	2.30	94.05	86.07
	Llama	N	Y	N	0.20	44.44	12.44	2.34	94.87	88.50
	Llama	Y	N	Y	0.62	80.89	41.78	1.30	91.84	73.75
	Llama	Y	Y	N	0.25	51.11	17.78	1.99	93.36	83.16
	Llama	Y	Y	Y	0.53	68.44	30.22	1.60	92.90	78.27
	Gemma	N	N	N	0.34	68.44	25.78	1.55	92.19	77.22
	Gemma	N	Y	N	0.32	67.56	24.89	1.62	92.60	79.23
	Gemma	Y	N	Y	0.48	82.67	39.56	1.24	91.07	70.98
	Gemma	Y	Y	N	0.33	68.89	28.44	1.62	92.19	77.59
	Gemma	Y	Y	Y	0.40	72.89	30.67	1.47	91.94	75.80

Table 7: Comprehensive results for All Datasets.

Model	CEFR-SP (Partial)	CEFR-SP (Whole)
CEFR-SP-estimator	0.937	0.933
README-estimator_EN	0.817	0.797
RSRS_BERT_L	0.535	0.515
RSRS_BERT_B	0.552	0.534
RSRS_XLMR_L	0.67	0.645
RSRS_XLMR_B	0.678	0.654
RSRS_mBERT	0.627	0.609
RSRS_mT5_L	0.417	0.4
RSRS_mT5_L	0.611	0.585
RSRS_mT5_L	0.507	0.478
FKGL	0.734	0.705
ARI	0.71	0.68
SL	0.659	0.627

Table 8: Comparing CEFR-SP-estimator, README-estimator, with unsupervised models and traditional metrics on CEFR-SP datasets. Performance measured with Spearman's Correlation ρ ($\uparrow$)

Model	A1	A2	B1	B2	C1	C2	Average F1
CEFR-SP-estimator	50.00	45.10	50.56	56.41	56.76	28.57	47.9
README-estimator	55.81	57.97	59.52	66.67	57.58	0.00	49.59

Table 9: Comparing CEFR-SP-estimator and README-estimator on README_EN dataset. Performance measured with F1.

-	A1	A2	B1	B2	C1	Total
English:						
Train	96	458	1277	206	10	2045
Val	26	133	303	47	3	512
French:						
Train	0	908	700	345	23	1976
Val	0	220	179	93	3	495
Russian:						
Train	428	772	368	72	0	1640
Val	117	182	86	26	0	411

Table 10: Distribution of sentence target levels in training and validation sets for Supervised models

Corpus	Source	Target	DP-Path	RRIS↑
CEFR-SP-Partial	C2	B1	[B2, B1]	25.0%
	C2	A2	[B2, B1, A2]	
	C2	A1	[B2, B1, A2, A1]	
CEFR-SP-Whole	C2	B1	[B2, B1]	25.0%
	C2	A2	[B2, B1, A2]	
	C2	A1	[B2, B1, A2, A1]	
README_EN	C2	B1	[B2, B1]	25.0%
	C2	A2	[B2, B1, A2]	
	C2	A1	[B2, B1, A2, A1]	
README_AR	C2	B1	[B2, B1]	28.57%
	C2	A2	[B2, B1, A2]	
README_FR	C2	A2	[B2, B1, A2]	22.2%
	C2	A1	[B2, B1, A2, A1]	
README_RU	C2	B1	[B2, B1]	27.27%
	C2	A2	[B2, B1, A2]	
	C2	A1	[B2, B1, A2, A1]	
	C1	A2	[B1, A2]	
	C1	A1	[B1, A2, A1]	
	B2	A1	[A2, A1]	
README_HI	C2	A2	[B1, A2]	42.11%
	C2	A1	[B1, A2, A1]	
	C1	A2	[B1, A2]	
	C1	A1	[B1, A2, A1]	
	B2	A1	[A1]	

Table 11: RRIS on Gemma. RRIS denotes the **R**eduction **R**atio of **I**nference **S**teps relative to the Sequential-steps strategy.