Simplify-Pro: A Two-level and Progressive LLM-based Framework for Auto Long Text Simplification

Peng Zhou[1] **Guangxin Li**[1,*] **Xiaoying Huang**[2] **Yiming Tang**[1]
[1] School of Computer Science and Technology, Xidian University, Xi'an, China
[2] School of Foreign Languages, Xidian University, Xi'an, China
kuangdaxufei@163.com, lgx@xidian.edu.cn

Abstract

Text simplification plays a vital role in natural language processing, yet auto long text simplification remains challenging due to the difficulty in the joint balancing of simplification efficiency and fine-grained quality requirements, such as fluency, grammatical correctness and semantic completeness. To address these challenges, we propose Simplify-Pro, a two-level and progressive LLM-based framework that establishes an effective paradigm for automatic long text simplification under diverse test scenarios. By integrating paragraph-level training, simplification generation, metric-assisted analysis and selective refinement into a unified multi-stage pipeline, our framework achieves superior performance across in-domain and out-of-domain simplification tasks, which matches or even outperforms advanced and proprietary LLMs. Furthermore, comprehensive experiments and qualitative analyses cover the simplification performance, generalization ability and the contribution of each individual stage, demonstrating the effectiveness, robustness and modular design advantages of Simplify-Pro.

1 Introduction

Text simplification has always been an important task in natural language processing (NLP), aiming to improve the comprehensibility of complex textual content while preserving meaning (Vásquez-Rodríguez et al., 2021; Ryan et al., 2023). Effective simplification enables broader access to information, which supports educational development, enhances accessibility for individuals with reading challenges and improves comprehension in specialized fields where technical terminology can be a barrier (Štajner, 2021; Sun et al., 2023a).

However, most existing studies have concentrated on lexical- and sentence-level simplification, leaving long text simplification comparatively underexplored despite its substantial practical importance in applications (Jiang et al., 2020; Scarton and Specia, 2018; Maddela et al., 2021; Agrawal and Carpuat, 2023). Moreover, current methods for long text simplification typically rely on large-scale and high-quality parallel data (Sun et al., 2023b). Meanwhile, errors in these methods become more frequent and obvious with increasing input length, including missing key information, semantic distortion and incoherent transitions (Ma et al., 2022; Wolska and Clausen, 2017), as shown in Figure 1. As a result, the proper balancing of simplification efficiency and fine-grained quality requirements such as fluency and semantic completeness remains challenging for complex long text.

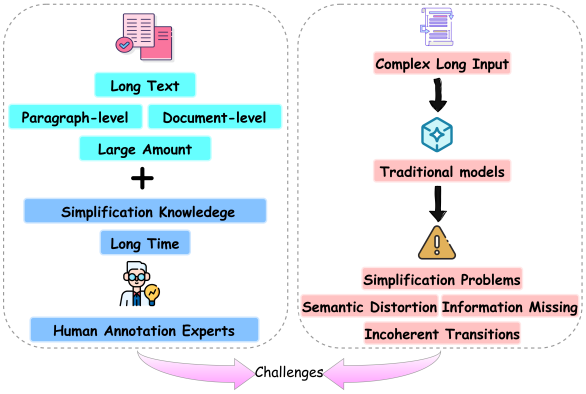

Figure 1: Challenges in automatic long text simplification. Existing approaches heavily rely on large-scale and human-annotated data, and suffer from obvious simplification problems with increasing input length.

In this paper, we introduce Simplify-Pro, a two-level and progressive LLM-based framework that can perform long text simplification at both the paragraph and document levels. The entire framework contains four stages: (1) Paragraph-level training: LLMs only trained on limited paragraph-level data serve as the core simplification component. (2) Simplification generation: based on the trained simplification model, paragraph-level inputs can be processed directly, and document-

*Corresponding author

Findings of the Association for Computational Linguistics: ACL 2026, pages 22096–22122

level text is segmented into a few text blocks, each simplified independently and then concatenated together. (3) Metric-assisted analysis: an evaluator is designed to conduct quantitative analysis that captures the structural and content transformations of generated simplifications. (4) Selective refinement: multiple candidate simplifications are paired with metric-assisted analyses and provided to a LLM editor, which then selects the most suitable simplification and performs lexical refinements. Our core contributions can be summarized as follows:

- We introduce a novel, two-level and progressive LLM-based framework, Simplify-Pro, which establishes an effective paradigm for long text simplification under diverse test scenarios.
- Our proposed framework achieves superior performance across in-domain and out-of-domain simplification tasks, which matches or even outperforms advanced and proprietary LLMs.
- Comprehensive experiments and analyses cover simplification performance, generalization ability and the necessity of each individual stage, providing systematic insights into the effectiveness and robustness of Simplify-Pro.

2 Related Work

Traditional text simplification. Most prior work has focused on lexical- and sentence-level simplification, where sequence-to-sequence models are used for complex-to-simple transformation (Nisioi et al., 2017; Zhang and Lapata, 2017; Zhao et al., 2018; Vu et al., 2018). For long text simplification, Laban et al. (2021) proposed a multi-paragraph level unsupervised method. Cripwell et al. (2023) implemented a plan-guided system where a planner predicts edits on the current sentence using document context. In addition, Laban et al. (2023) fine-tuned the bart model on a large-scale dataset from Wikipedia to generate long text simplification.

LLM-based text simplification. Recent studies have explored some directions to investigate the capabilities of large language models (LLMs) for text simplification. Kew et al. (2023) constructed benchmarks and compared prompting strategies to assess LLMs' performance for sentence simplification. Also, Agrawal and Carpuat (2024) further observed that prompted LLMs simplify adequately but remain inferior to supervised systems in terms of fluency. And others have developed LLM-driven systems customized for specific user groups, such as language learners and readers with limited literacy skills (Shaib et al., 2023; Athugodage et al., 2024; Mo and Hu, 2024; Färber et al., 2025).

3 Methodology

3.1 Framework Overall

As illustrated in Figure 2, the overall framework follows a multi-stage pipeline that integrates paragraph-level training (Section 3.2), simplification generation (Section 3.3), metric-assisted analysis (Section 3.4) and selective refinement (Section 3.5). Each stage of Simplify-Pro is modular and necessary, which enables progressive improvements in simplification quality.

3.2 Paragraph-level Training

The core components of our framework are LLMs which can directly handle paragraph simplification. Here, we only use limited paragraph-level data and choose the Group Relative Policy Optimization (GRPO) algorithm (Shao et al., 2024) to train the simplification model. Given a simplification question-answer pair (q, a), the behavior policy $\pi_{\theta_{\text{old}}}$ samples a group of G individual responses $\{o_i\}_{i=1}^{G}$ at each training step. Then GRPO maximizes the clipped objective as follows:

$$J_{GRPO}(\theta) = \mathrm{E}_{(q,a)\sim D,\ \{o_i\}_{i=1}^{G}\sim\pi_{\theta_{old}}(\cdot|q)} \Bigg[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\min\Big(r_{i,t}(\theta)\hat{A}_{i,t}, \mathrm{clip}(r_{i,t}(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_{i,t}\Big) - \beta\, D_{\mathrm{KL}}(\pi_\theta \,\|\, \pi_{ref}) \Bigg] \tag{1}$$

where ϵ constrains the allowable range of policy updates and β avoids diverging excessively from the reference policy π_{ref}. The advantage $\hat{A}_{i,t}$ is computed using the normalized reward within the group $\{R_i\}_{i=1}^{G}$.

Then an instruction-guided prompt is adopted to encourage LLMs to generate structured outputs and detailed reasoning steps, as shown in Appendix D. We also design a synergistic reward combining both strict format verification and multi-dimensional evaluation of the simplification quality.

For strict format verification, outputs satisfying the predefined formatting rules are rewarded positively. Otherwise, a $penalty$ is applied to structural violations. Since the training prompt doesn't restrict specific editing operations, an overly concise

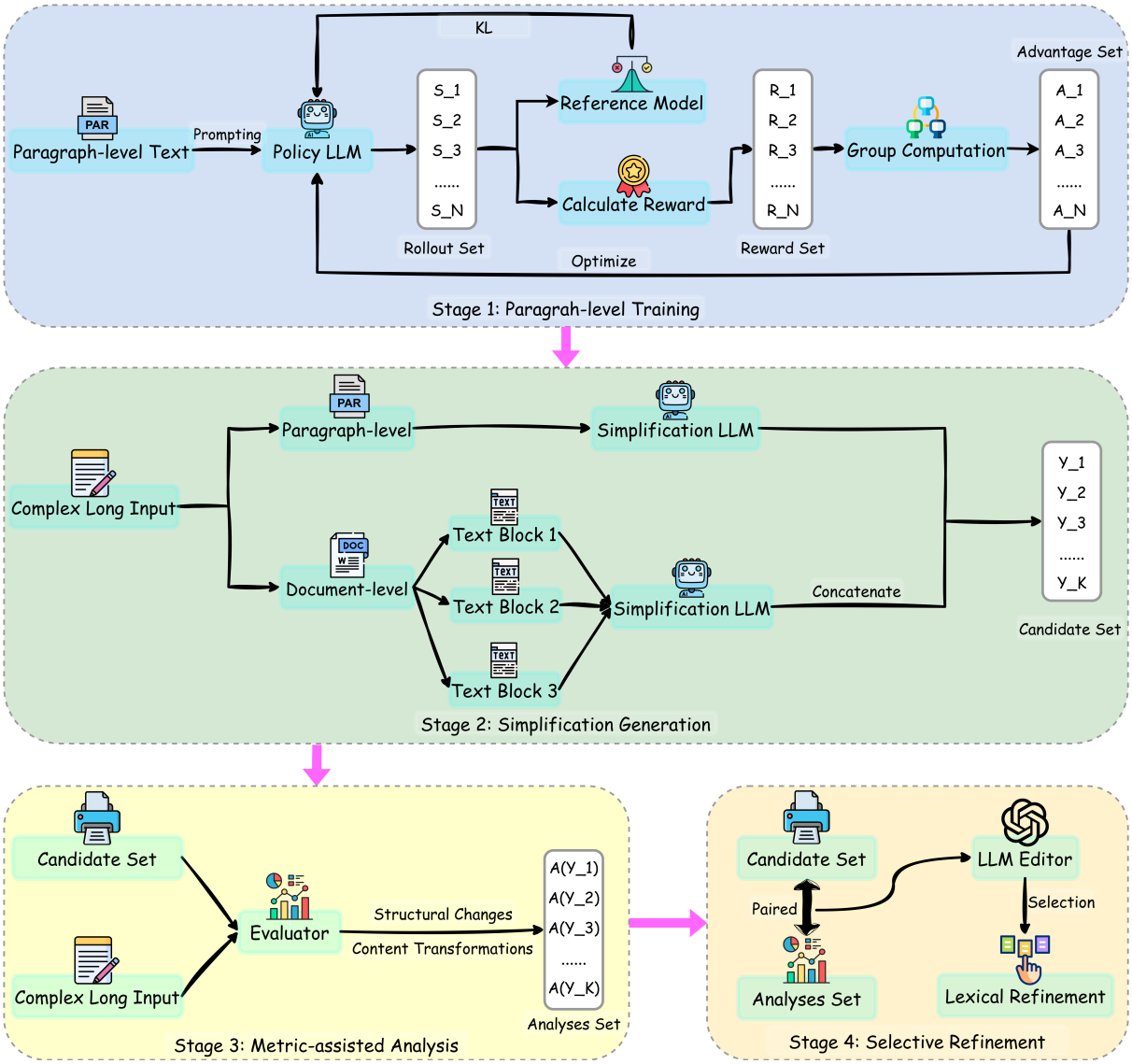

Figure 2: An overview of Simplify-Pro, to perform paragraph- and document-level simplifications progressively.

reasoning process may limit the model's ability to explore diverse simplification strategies. Also, excessively long reasoning may produce many meaningless intermediate steps which can increase optimization noise during training. Thus, a dynamic reasoning reward R_{CoT} is introduced to regulate the length of the reasoning process L_{CoT} adaptively. The punishment increases proportionally with the degree of deviation when L_{CoT} falls outside the predefined range ($[L_{lower}, L_{upper}]$).

$$R_{CoT} = \min\left(\frac{L_{CoT} - L_{lower}}{L_{lower}} \cdot F_{lower},\ 0\right) + \max\left(\frac{L_{CoT} - L_{upper}}{L_{upper}} \cdot F_{upper},\ 0\right) \quad (2)$$

where L_{lower} and L_{upper} respectively denote the expected minimal and maximum length of L_{CoT}. F_{lower} and F_{upper} are scaling factors that control the strength of penalties.

Given that text simplification is an open-ended task, discrete rewards are inadequate for providing fine-grained feedback during training. Therefore, we consider the following three aspects for more comprehensively evaluating simplification quality.

1. **Lexical Edit Operations** aim to assess how effectively a simplification system performs lexical modifications to reduce textual complexity. Here, SARI is chosen to measure the F1 scores of n-grams that are added, deleted or retained (Xu et al., 2016).
2. **Semantic Similarity** measures how well the meaning of original text is preserved in the final simplification. We use BERTScore which leverages contextualized representations to capture semantic similarity (Zhang et al., 2020).
3. **Human Preference Alignment** effectively evaluates whether outputs align with human judgments of acceptable simplification. Therefore,

LENS is used to reveal human preference alignment by producing a score ranging from 0 to 100 (Maddela et al., 2023).

Based on our exploration, the simplification quality reward $R_{quality}$ is formulated as follows:

$$R_{quality} = \begin{cases} R_{lex}, & \text{if RS} \\ R_{lex} + R_{hpa}, & \text{if RD} \\ R_{lex} + R_{sem} + R_{hpa}, & \text{if RT} \end{cases} \quad (3)$$

where R_{lex}, R_{sem} and R_{hpa} denote the reward components corresponding to lexical edit operations, semantic similarity and human preference alignment, respectively. RS, RD and RT control the combination of the simplification quality reward.

Finally, the total reward R is defined as follows:

$$R = \begin{cases} R_{CoT} + R_{quality} + 1, & \text{if correct format} \\ penalty, & \text{otherwise} \end{cases} \quad (4)$$

3.3 Simplification Generation

Paragraph text can be complex enough to involve meaningful structural and content simplifications (Agrawal et al., 2021), while remaining relatively short to be processed. Thus, our training models can be directly used to simplify paragraph-level inputs in a stable and predictable manner.

For document-level simplification, we adopt a dynamic, semantics-aware segmentation strategy to decompose long inputs into a few coherent text blocks. Specifically, each sentence of the complex text is encoded by a pretrained sentence embedding model to obtain a vector representation.

$$V_i = Encoder(S_i), i = 1, 2, 3, \ldots, M \quad (5)$$

where S_i and M denote the i-th sentence and the total number of sentences in the input, respectively.

Then semantic similarity Sim_i between adjacent sentences is computed to measure local semantic continuity in the original text. A significant drop in Sim_i indicates a potential semantic or topical shift and can be treated as a potential boundary.

$$Sim_i = \frac{V_i^\top \cdot V_{i+1}}{\|V_i\| \, \|V_{i+1}\|}, i = 1, 2, 3, \ldots, M-1 \quad (6)$$

Given the similarity sequence $\{Sim_i\}_{i=1}^{M-1}$, we sort all values and select initial boundaries whose similarity scores fall below a low threshold τ. To locate more reliable semantic boundaries, a greedy merging strategy with soft length constraints is applied. Starting from the initial segmentation, adjacent blocks are iteratively merged in order to keep the token length of each block falling within a target interval $\left[0.75 \cdot \frac{L_{src}}{N}, 1.25 \cdot \frac{L_{src}}{N}\right]$ whenever possible. This dynamic merging step yields a final set of text blocks that are semantically coherent and suitable for independent simplification. N can be computed as follows:

$$N = \max\left(\left\lceil \ln\left(\frac{L_{src} + L_{pre}}{|L_{src} - L_{pre}|}\right)\right\rceil, \left\lceil \frac{L_{src}}{L_{pre}} \right\rceil\right) \quad (7)$$

where L_{src} denotes the token length of the input text, and L_{pre} is a predefined value.

Then, each block is simplified using the trained paragraph-level model and concatenated following their original order to reconstruct the final simplification. This block-by-block simplification strategy prevents unexpected performance drops caused by overly long inputs.

3.4 Metric-assisted Analysis

In this stage, we design a metric-assisted evaluator which captures structural and content transformations focusing on generated simplifications. In particular, structural changes are assessed through the proportions of additions, deletions and copies ($P_{add}, P_{del}, P_{copy}$), which reflect how the simplified text differs from the original input in terms of editing operations.

$$|\mathcal{E}(X \to Y)| = N_{add} + N_{del} + N_{copy} \quad (8)$$

$$P_{edit} = \frac{N_{edit}}{|\mathcal{E}(X \to Y)|}, edit \in \{add, del, copy\} \quad (9)$$

Here, X and Y denote the original long text and its simplification. N_{add}, N_{del} and N_{copy} count the corresponding editing operations that transform X into Y at the token level, respectively.

Content transformations are analyzed using a combination of Levenshtein similarity and lexical complexity. $LevSim$ measures the extent of editing required to transform the original text into the simplification, while $LexC$ is calculated to measure whether simplification leads to the use of simpler and more accessible vocabulary.

$$LevSim(X, Y) = 1 - \frac{d_{lev}(X, Y)}{\max(|X|, |Y|)} \quad (10)$$

$$LexC(Y) = -\frac{1}{|Y|} \sum_{w \in Y} \log\left(f(w) + \delta\right) \quad (11)$$

where $|X|$ and $|Y|$ denote the length of X and Y. The function $d_{lev}(X, Y)$ represents the Levenshtein distance. $f(w)$ denotes the general usage frequency of a word estimated from a large reference corpus, and δ is a small positive constant introduced for numerical stability.

3.5 Selective Refinement

Due to the training on paragraph-level data consisting of multiple sentences, our models naturally acquire sentence-level simplification capabilities, including syntactic rewritin and structural reorganization. However, this training process does not explicitly target fine-grained lexical choice, leaving room for improvement when structure and content organization are already appropriate. Therefore, the final stage is dedicated to lexical refinement for further enhancing simplification quality.

In practice, human experts typically address this aspect by first producing several simplified drafts, comparing them from multiple perspectives and then selecting a promising version for careful lexical polishing (Zhong et al., 2020). Inspired by this, we employ a selective refinement strategy. For a given complex input X, a set of candidate outputs $\{Y_i\}_{i=1}^{K}$ are produced during the stage of simplification generation. Then each candidate Y_i is paired with the corresponding metric-assisted analysis $A(Y_i)$, providing an explicit and comparable characterization of simplification behaviors. Based on the set $\{(Y_i, A(Y_i))\}_{i=1}^{K}$, a LLM editor is prompted to select the most suitable simplification Y^*, and subsequently apply lexical refinement while preserving the overall structure and content. This strategy mirrors the iterative workflow of human experts, enabling controlled lexical improvements and more stable simplification.

4 Experiments

4.1 Experimental Setup

Datasets. Our primary experiments focus on long text simplification in the Medical and Wikipedia domains. For paragraph-level training and evaluations, we use the official train/val/test splits of a medical dataset Cochrane (Devaraj et al., 2021) and collect 4434 parallel samples from the D-Wikipedia dataset (Sun et al., 2021). For document-level evaluations, we select long text exceeding 1000 tokens from another medical dataset Cell (Guo et al., 2024) and the Wikipedia-based dataset Swipe (Laban et al., 2023). Meanwhile, generalization capabilities are mainly evaluated on one out-of-domain document-level dataset, OneStopEnglish (News), which contains news articles written at three different reading levels (Vajjala and Lučić, 2018).

Baselines. Our baselines cover a diverse set of advanced open-source models, including Qwen2.5-72B-Instruct (Team, 2024), Llama3.3-70B-Instruct (Grattafiori et al., 2024) and DeepSeek-V3.2 (Liu et al., 2025), alongside leading proprietary models, namely Qwen3-Max (Yang et al., 2025), GPT-5.2, Claude-Sonnet-4.5 and Gemini-3-Pro-Preview. We further include two traditional baselines for comparison: Keep-It-Simple, a multi-paragraph-level unsupervised method for text simplification (Laban et al., 2021), and Bart-Large-Wiki, a model finetuned on Wikipedia-based simplification data (Laban et al., 2023). However, due to their design constraints, the two methods above struggle with handling document-level inputs longer than 1,000 tokens. So their results are reported only in the paragraph-level simplification setting.

Metrics. To provide a more comprehensive assessment of simplification quality, we adopt five complementary metrics, including SARI (Xu et al., 2016), D-SARI (Sun et al., 2021), FKGL (Kincaid et al., 1975), BERTScore (Zhang et al., 2020) and LENS (Maddela et al., 2023).

The datasets statistics, baseline descriptions, metrics introduction and implementation details are provided in Appendix A – Appendix D.

4.2 Main Results

In-Domain performance. Table 1 and Table 2 report comparative results on both the Medical and Wikipedia domains at paragraph and document levels. Traditional baselines lag far behind on paragraph-level simplification, indicating clear limitations in handling complex long text. By comparison, strong open-source and proprietary LLM baselines are more competitive. For example, Llama3.3-70B-Instruct yields relatively high LENS scores, while Gemini-3-Pro-Preview attains strong SARI and D-SARI values. However, these methods often reveal noticeable trade-offs between structural simplification and fine-grained quality control. In contrast, Simplify-Pro outperforms all baseline groups across most metrics. At the paragraph level, Simplify-Pro-RS reaches the best SARI scores (44.99/48.80) with superior D-SARI and BERTScore values. Simplify-Pro-RD shows clear advantages in readability and overall simplification quality, achieving the best FKGL

Method	Cochrane(Medical)					D-Wikipedia(Wikipedia)				
	SARI↑	D-SARI↑	FKGL↓	BERT↑	LENS↑	SARI↑	D-SARI↑	FKGL↓	BERT↑	LENS↑
Traditional Baseline										
Keep-It-Simple	35.09	27.13	10.30	13.89	65.09	37.67	29.41	8.19	20.70	65.76
Bart-Large-Wiki	34.62	27.74	10.47	14.67	63.83	40.52	31.77	5.46	24.34	67.41
Strong Open-source LLMs Baseline										
Qwen2.5-72B-Instruct*	41.61	29.59	8.66	20.97	71.17	43.77	32.74	7.11	27.27	74.99
Llama3.3-70B-Instruct*	42.38	29.26	9.12	21.87	79.92	45.49	32.36	7.13	26.94	79.94
DeepSeek-V3.2†	42.70	30.42	7.67	23.25	76.65	45.04	33.30	5.59	25.69	76.28
Strong Closed-source LLMs Baseline										
Qwen3-Max†	41.43	29.09	8.92	19.84	75.58	44.49	31.31	7.13	24.71	74.81
Gemini-3-Pro-Preview†	42.90	30.76	6.87	21.36	77.96	46.06	33.95	5.51	25.32	72.09
GPT-5.2†	41.94	29.80	8.64	20.56	80.72	45.21	32.72	6.70	26.05	77.29
Claude-Sonnet-4.5†	42.09	31.43	8.84	22.43	74.18	43.95	32.20	7.36	26.79	79.28
Backbone Baseline										
Qwen2.5-14B-Instruct	40.28	27.43	8.70	20.15	72.37	42.20	30.32	6.99	23.02	73.49
Qwen2.5-14B-Instruct-SFT	41.93	28.28	9.19	25.90	73.33	44.07	31.15	6.17	29.16	74.95
Ours										
Simplify-Pro-RS	**44.99**	32.89	9.99	25.47	77.15	**48.80**	**39.70**	3.33	31.25	77.22
Simplify-Pro-RD	43.42	30.00	**6.29**	21.36	**90.14**	45.81	33.06	**2.15**	26.52	**90.74**
Simplify-Pro-RT	44.64	**36.19**	8.74	**28.49**	87.25	47.59	37.23	3.28	**34.64**	84.56

Table 1: Performance comparison on in-domain and paragraph-level simplification. * indicates the use of in-context learning (ICL) during inference, and † denotes the use of both ICL and chain-of-thought (CoT) prompting. For our framework, RS, RD and RT represent using different reward strategies in the paragraph-level training stage. The best and second performance are in bold and underlined, respectively.

Method	Cell(Medical)					Swipe(Wikipedia)				
	SARI↑	D-SARI↑	FKGL↓	BERT↑	LENS↑	SARI↑	D-SARI↑	FKGL↓	BERT↑	LENS↑
Strong Open-source LLMs Baseline										
Qwen2.5-72B-Instruct*	43.78	36.74	8.69	15.88	76.15	42.01	34.12	7.38	20.45	73.57
Llama3.3-70B-Instruct*	44.80	37.93	10.11	16.74	82.71	43.53	34.94	8.48	21.16	78.33
DeepSeek-V3.2†	44.84	37.92	7.29	18.38	79.43	43.99	35.21	5.97	21.65	74.70
Strong Closed-source LLMs Baseline										
Qwen3-Max†	43.81	36.68	8.49	16.06	75.74	42.32	33.23	7.45	20.06	72.71
Gemini-3-Pro-Preview†	44.61	37.93	7.56	17.07	76.29	44.19	35.91	6.31	21.37	75.51
GPT-5.2†	45.13	38.01	9.04	15.75	73.76	43.06	34.11	7.19	20.81	80.02
Claude-Sonnet-4.5†	44.47	37.43	9.71	16.32	77.73	42.89	34.53	7.78	22.04	76.07
Backbone Baseline										
Qwen2.5-14B-Instruct	42.10	36.08	7.73	15.11	74.08	41.40	32.99	6.46	19.44	73.60
Qwen2.5-14B-Instruct-SFT	43.18	36.86	9.58	17.85	73.83	42.64	33.37	5.85	21.80	72.66
Ours										
Simplify-Pro-RS	**47.71**	39.72	10.16	21.10	77.91	**47.47**	**40.77**	3.76	26.58	77.12
Simplify-Pro-RD	44.81	38.68	**6.68**	17.19	**89.03**	44.59	35.53	**2.26**	21.32	**89.29**
Simplify-Pro-RT	46.69	**39.95**	9.18	**22.68**	86.71	45.83	37.57	3.38	**28.61**	83.80

Table 2: Performance comparison on in-domain and document-level simplification.

(6.29/2.15) and LENS scores (90.14/90.74). Meanwhile, Simplify-Pro-RT further boosts fine-grained structure and semantic preservation, with top or near-top D-SARI (36.19/37.23) and BERTScore (28.49/34.64). Similar trends persist at the document level, where Simplify-Pro-RS and Simplify-Pro-RT attain the strongest or near-strongest performance across SARI, D-SARI and BERTScore metrics. Also, Simplify-Pro-RD remains the lowest FKGL (6.68/2.26) and highest LENS (89.03/89.29) across both domains. Overall, these results demonstrate the effectiveness of our approach even in comparison with powerful proprietary LLMs, highlighting the robustness of Simplify-Pro for complex long text simplification.

Out-of-Domain performance. Table 3 presents the document-level simplification results on the out-of-domain dataset. Strong open-source and closed-source LLM baselines demonstrate strong results. For instance, Claude-Sonnet-4.5 attains the high-

Method	OneStopEnglish (News)		
	D-SARI↑	FKGL↓	LENS↑
Strong Open-source LLMs Baseline			
Qwen2.5-72B-Instruct*	28.48	7.54	72.71
Llama3.3-70B-Instruct*	28.85	7.68	77.48
DeepSeek-V3.2†	29.54	5.77	75.47
Strong Closed-source LLMs Baseline			
Qwen3-Max†	28.10	6.67	75.28
Gemini-3-Pro-Preview†	29.31	6.44	77.05
GPT-5.2†	29.64	6.97	74.15
Claude-Sonnet-4.5†	**29.68**	6.87	78.54
Ours			
Simplify-Pro-RS	27.70	7.43	78.65
Simplify-Pro-RD	28.54	**4.01**	**87.23**
Simplify-Pro-RT	28.96	6.36	84.73

Table 3: Performance comparison on out-of-domain and document-level simplification. For our framework, we directly use simplification models trained only on paragraph-level data from the Medical domain.

est D-SARI (29.68), while DeepSeek achieves the second-lowest FKGL (5.77). However, their performance varies substantially across different metrics. By contrast, our framework exhibits more stable and transferable simplification behavior. In particular, Simplify-Pro-RT achieves competitive D-SARI performance, the third-lowest FKGL (6.36) and the second-best LENS score (84.73), reflecting more effective simplification in the unfamiliar domain. Meanwhile, Simplify-Pro-RD yields the lowest FKGL score (4.01) and the highest LENS score (87.23), indicating improved readability and overall simplification quality. Notably, these gains are achieved without any domain-specific training, highlighting the robustness and cross-domain generalization capability of our framework.

4.3 Performance Across Models

To further validate the generalization ability of Simplify-Pro, we investigate whether the framework remains effective when instantiated with different LLMs. Table 4 reports the performance comparison on document-level simplification in the Medical domain. For all three framework variants, Simplify-Pro with Qwen2.5-14B-Instruct yields relatively higher scores than Llama3.1-8B-Instruct across most metrics. The former improves SARI by about 1.12–1.50 and D-SARI by 0.46–0.96, with additional gains in BERTScore (up to +2.32) and LENS (up to +2.49). A similar trend remains at the paragraph-level, where we also provide results in Appendix G. The stronger performance can be largely attributed to Qwen's larger parameter scale, which provides greater modeling capacity for learning fine-grained edit operations while balancing semantic preservation and overall quality. Notably, Simplify-Pro with Llama sometimes achieves better FKGL or SARI scores, and exhibits performance which still outperforms most baseline methods mentioned in Table 1 and Table 2. These results suggest that the effectiveness of Simplify-Pro is not specific to certain model series, providing support for the soundness of its design.

Method	SARI	D-SARI	FKGL	BERT	LENS
Qwen2.5-14B-Instruct for paragraph-level training					
Simplify-Pro-RS	47.71	39.72	10.16	21.10	77.91
Simplify-Pro-RD	44.81	38.68	6.68	17.19	89.03
Simplify-Pro-RT	46.69	39.95	9.18	22.68	86.71
Llama3.1-8B-Instruct for paragraph-level training					
Simplify-Pro-RS	46.21	38.76	11.66	19.81	75.42
Simplify-Pro-RD	45.17	38.22	6.49	14.87	87.82
Simplify-Pro-RT	45.57	39.29	9.08	21.50	84.30

Table 4: Performance comparison on document-level simplification in the Medical domain.

Method	SARI	D-SARI	FKGL	BERT	LENS
Paragraph-level Simplification					
Simplify-Pro-RT	**44.64**	**36.19**	**8.74**	28.49	**87.25**
w/o MA	44.03	35.39	8.90	28.22	86.68
w/o MA&SR	43.68	35.66	9.35	**28.70**	86.16
Document-level Simplification					
Simplify-Pro-RT	**46.69**	**39.95**	9.18	**22.68**	**86.71**
w/o MA	46.18	39.58	9.04	22.41	86.37
w/o MA&SR	45.66	39.24	**8.95**	21.50	85.79

Table 5: Ablation study on Simplify-Pro-RT's simplification performance in the Medical domain. MA and SR refer to the stage of quality analysis and selective refinement, respectively. The best results are in bold.

4.4 Ablation Study

To systematically assess the necessity of different stages, we conduct ablation studies focusing on Simplify-Pro-RT, with results reported in Table 5 for both paragraph-level and document-level simplification. Removing the metric-assisted analysis (MA) stage consistently leads to performance degradation, particularly on lexical-edit metrics. In the Medical domain, SARI drops from 44.64 to 44.03 and D-SARI decreases from 36.19 to 35.39 at the paragraph level, with similar declines observed at the document level. These trends indicate that MA provides essential guidance for comparing candidates and preserving editing simplification gains. Further removing the selective refinement (SR) stage results in additional degradation, most notably on semantic (BERTScore) and quality-

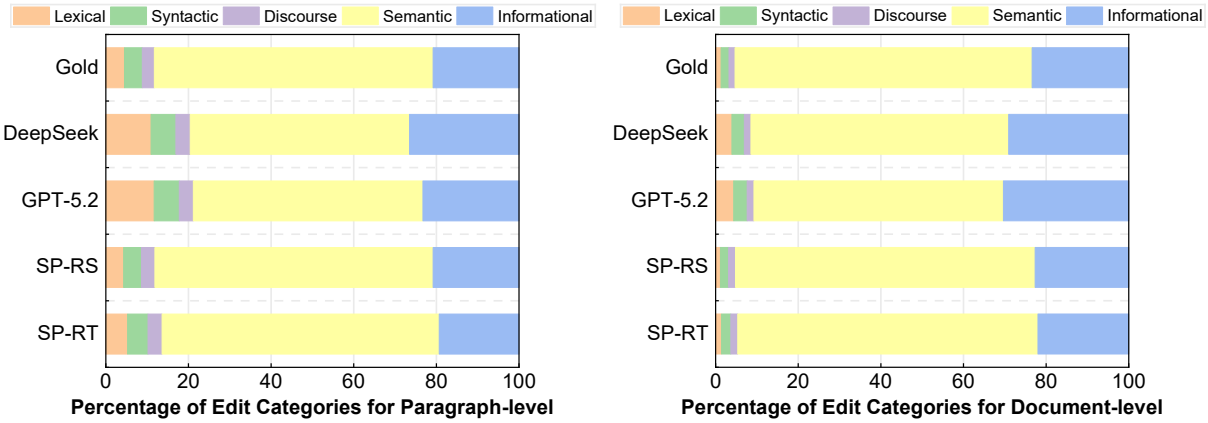

Figure 3: Distribution of five edit categories for paragraph- and document-level simplifications in the Medical domain. Gold, SP-RS and SP-RT denote reference simplifications, Simplify-Pro-RS and Simplify-Pro-RT, respectively.

Method	Simplicity	Readability	Meaning Preservation	Preference (Count / 80)
DeepSeek-v3.2	3.93	3.55	**4.23**	13
Gemini-3-Pro-Preview	<u>4.25</u>	3.98	3.85	<u>16</u>
GPT-5.2	3.68	<u>4.03</u>	3.73	7
Simplify-Pro-RS	4.20	3.78	3.95	13
Simplify-Pro-RD	4.03	**4.15**	3.48	12
Simplify-Pro-RT	**4.58**	3.83	<u>4.08</u>	**19**

Table 6: Results of human evaluation on 80 randomly sampled test instances. The best and second performance are in bold and underlined, respectively. For Simplicity, Readability, and Meaning Preservation, we report the average scores across all annotators. Preference indicates the number of times a system output is selected as the overall preferred version across the 80 evaluated instances.

oriented (LENS) metrics (22.68 → 21.50, 87.25 → 86.16). While variants without MA or SR occasionally yield slightly better FKGL or BERTScore (8.95/28.70), the complete framework achieves the most balanced performance across diverse scenarios. These results confirm that both MA and SR are essential stages of the framework, jointly contributing to stable and high-quality long text simplification. We also provide ablation results in the Wikipedia domain in Appendix G. More discussions of different stages and simplification case studies are provided in Appendix F – Appendix K.

4.5 Simplification Analysis

To better assess the effectiveness of our framework in long text simplification, we group all applied simplification operations into five categories (Lexical, Syntactic, Discourse, Semantic and Informational) and analyze their relative distributions. As shown in Figure 3, the edit distributions produced by Simplify-Pro variants are substantially closer to the gold human-annotated reference across both paragraph-level and document-level simplifications in the Medical domain. This alignment suggests that Simplify-Pro adopts more appropriate and human-like simplification strategies. In contrast, strong baseline LLMs such as DeepSeek and GPT exhibit noticeably different distribution patterns, particularly in lexical, semantic and informational edits. Moreover, it can be observed that semantic edits account for the highest proportion of overall operations, which indicates that effective long text simplification is primarily driven by semantic-level content selection and transformation.

4.6 Human Evaluation

Automatic evaluations are not always indicative of performance. Therefore, a human evaluation is conducted on 80 simplification instances. To ensure diversity and reduce domain bias, we randomly sample 20 instances from each of four datasets used in the main experiments (Table 1 and Table 2). Each simplification is rated on a 1–5 scale along three dimensions: Simplicity, Readability, and Meaning Preservation, where higher scores indicate better quality. In addition, annotators are asked to select the overall preferred output for each instance.

As shown in Table 6, the human evaluation results reveal clear trade-offs among systems. Simplify-Pro-RT achieves the highest Simplicity score (4.58) and overall preference count (19 out of 80), outperforming the second-best Gemini-3-pro-preview (4.25, 16 out of 80). This indicates that annotators favor Simplify-Pro-RT when consider-

ing simplification quality holistically. Importantly, Simplify-Pro-RT also attains the second-highest Preservation score (4.08), close to the best Preservation model (DeepSeek-v3.2, 4.23), suggesting that its better simplification does not come at the cost of major meaning distortion.

5 Conclusion and Future Work

In this paper, we propose Simplify-Pro, a two-level and progressive LLM-based framework for effective auto long text simplification. By integrating paragraph-level training, simplification generation, metric-assisted analysis and selective refinement into a multi-stage pipeline, our framework can rival or even surpass much larger and advanced proprietary LLMs on both in-domain and out-of-domain simplification tasks. Further experiments and analyses highlight our framework's generalization ability and the necessity of each individual stage. Future efforts aim to enhance our research by adapting the framework to multilingual and low-resource text simplification scenarios.

Limitations

Future efforts to improve this work should address the following limitations through two primary directions:

- Firstly, the impact of alternative model architectures and different sizes of model parameters could be further explored, while our proposed framework achieves strong performance with Qwen2.5-14B-Instruct as the core simplification models. Evaluating a wider range of models with varying parameter sizes could provide deeper insights into how model capacity influences simplification performance.
- Secondly, our framework is primarily evaluated in the English domain despite achieving strong performance on text simplification tasks. Consequently, the effectiveness of the proposed framework across a broader range of languages has not been systematically assessed, particularly in low-resource language settings.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Sweta Agrawal and Marine Carpuat. 2022. An imitation learning curriculum for text editing with non-autoregressive models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7550–7563.

Sweta Agrawal and Marine Carpuat. 2023. Controlling pre-trained language models for grade-specific text simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12807–12819.

Sweta Agrawal and Marine Carpuat. 2024. Do text simplification systems preserve meaning? a human evaluation via reading comprehension. *Transactions of the Association for Computational Linguistics*, 12.

Sweta Agrawal, Weijia Xu, and Marine Carpuat. 2021. A non-autoregressive edit-based approach to controllable text simplification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3757–3769.

F Alva-Manchego, L Martin, C Scarton, and L Specia. 2019a. Easse: easier automatic sentence simplification evaluation. 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): System Demonstrations*, pages 49–54. Association for Computational Linguistics.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2019b. Cross-sentence transformations in text simplification. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 181–184.

Mark Athugodage, Olga Mitrofanove, and Vadim Gudkov. 2024. Transfer learning for russian legal text simplification. In *Proceedings of the 3rd Workshop on Tools and Resources for People with REAding DIfficulties (READI)@ LREC-COLING 2024*, pages 59–69.

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.

Ashwin Devaraj, Iain Marshall, Byron C Wallace, and Junyi Jessy Li. 2021. Paragraph-level simplification of medical texts. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 4972–4984.

Ashwin Devaraj, William Sheffield, Byron C Wallace, and Junyi Jessy Li. 2022. Evaluating factuality in text simplification. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2022, page 7331.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu,

Baobao Chang, and 1 others. 2024. A survey on in-context learning. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 1107–1128.

Michael Färber, Parisa Aghdam, Kyuri Im, Mario Tawfelis, and Hardik Ghoshal. 2025. Simplifymytext: An llm-based system for inclusive plain language text simplification. In *European Conference on Information Retrieval*, pages 418–424. Springer.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yue Guo, Wei Qiu, Gondy Leroy, Sheng Wang, and Trevor Cohen. 2024. Retrieval augmentation of large language models for lay language generation. *Journal of Biomedical Informatics*, 149:104580.

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.

Tannon Kew, Alison Chi, Laura Vásquez-Rodríguez, Sweta Agrawal, Dennis Aumiller, Fernando Alva-Manchego, and Matthew Shardlow. 2023. Bless: Benchmarking large language models on sentence simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13291–13309.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. *Institute for Simulation and Training*, pages 1–49.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626.

Philippe Laban, Tobias Schnabel, Paul Bennett, and Marti A Hearst. 2021. Keep it simple: Unsupervised simplification of multi-paragraph text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6365–6378.

Philippe Laban, Jesse Vig, Wojciech Kryściński, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. 2023. Swipe: A dataset for document-level simplification of wikipedia pages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10674–10695.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, and 1 others. 2025. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.

Xinyu Lu, Jipeng Qiang, Yun Li, Yunhao Yuan, and Yi Zhu. 2021. An unsupervised method for building sentence simplification corpora in multiple languages. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 227–237.

Yuan Ma, Sandaru Seneviratne, and Elena Daskalaki. 2022. Improving text simplification with factuality error detection. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 173–178.

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*, pages 3536–3553.

Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2023. Lens: A learnable evaluation metric for text simplification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16383–16408.

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.

Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 85–91.

Gustavo Paetzold. 2015. Reliable lexical simplification for non-native speakers. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 9–16.

Gustavo H Paetzold and Lucia Specia. 2016. Unsupervised lexical simplification for non-native speakers. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 3761–3767.

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

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.

Chantal Shaib, Millicent Li, Sebastian Joseph, Iain Marshall, Junyi Jessy Li, and Byron C Wallace. 2023. Summarizing, simplifying, and synthesizing medical evidence using gpt-3 (with varying success). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1387–1407.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Zhang, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297.

Sanja Štajner. 2021. Automatic text simplification for social good: Progress and challenges. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2637–2652.

Renliang Sun, Hanqi Jin, and Xiaojun Wan. 2021. Document-level text simplification: Dataset, criteria and baseline. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7997–8013.

Renliang Sun, Wei Xu, and Xiaojun Wan. 2023a. Teaching the pre-trained model to generate simple texts for text simplification. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9345–9355.

Renliang Sun, Zhixian Yang, and Xiaojun Wan. 2023b. Exploiting summarization data to help text simplification. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 39–51.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

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.

Laura Vásquez-Rodríguez, Matthew Shardlow, Piotr Przybyła, and Sophia Ananiadou. 2021. Investigating text simplification evaluation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 876–882.

Tu Vu, Baotian Hu, Tsendsuren Munkhdalai, and Hong Yu. 2018. Sentence simplification with memory-augmented neural networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 79–85.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Magdalena Wolska and Yulia Clausen. 2017. Simplifying metaphorical language for young readers: A corpus study on news text. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 313–318.

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.

Daiki Yanamoto, Tomoki Ikawa, Tomoyuki Kajiwara, Takashi Ninomiya, Satoru Uchida, and Yuki Arase. 2022. Controllable text simplification with deep reinforcement learning. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 398–404.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, and 1 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

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*.

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.

Sanqiang Zhao, Rui Meng, Daqing He, Andi Saptono, and Bambang Parmanto. 2018. Integrating transformer and paraphrase rules for sentence simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3164–3173.

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.

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.

A Datasets

Datasets	Level	Domain	Count	Len	Lang
Cochrane	Par	Med	480	478	En
D_Wikipedia	Par	Wiki	434	385	En
Cell	Doc	Med	352	1145	En
Swipe	Doc	Wiki	252	1095	En
OneStopEnglish	Doc	News	189	1033	En

Table 7: The details of all datasets for evaluations. Level indicates the simplification level, where Par and Doc denote paragraph-level and document-level, respectively. Domain specifies the source domain, including Med (Medical), Wiki (Wikipedia) and News. Count denotes the size of the corresponding test set, and Len is the average input token length. Lang indicates the corresponding language, where En denotes English.

Table 7 summarizes the datasets used for evaluation in this work. For all datasets, we apply a unified text cleaning procedure prior to evaluation to reduce potential noise and ensure data consistency. This step removes invisible or non-printable characters (e.g., zero-width spaces and Unicode control symbols) and redundant whitespace.

B Baseline

We compare our approach with a diverse set of strong baseline models, including both open-source and proprietary large language models.

- **Qwen2.5-72B-Instruct**: A large-parameter instruction-tuned language model developed by Alibaba. It demonstrates strong zero-shot generation and reasoning capabilities across a wide range of natural language tasks, with notable performance in code generation, mathematical problem-solving, and multi-turn dialogue.
- **Llama3.3-70B-Instruct**: An instruction-tuned variant of the Llama-3 model family released by Meta. Trained on high-quality curated data, it is widely adopted as a competitive open-source baseline, exhibiting robust performance in general understanding and instruction-following tasks.
- **DeepSeek-V3.2**: A large open-source language model optimized for complex reasoning and long-context understanding. It shows strong performance in scenarios requiring logical inference, multi-document question answering, and extended text generation, supporting an exceptionally long context window.
- **Qwen3-Max**: A proprietary large language model by Alibaba, designed to support advanced reasoning and long-context generation. It excels in knowledge-intensive question answering, comprehensive analysis, and creative writing tasks.
- **GPT-5.2**: A proprietary large language model developed by OpenAI, exhibiting strong zero-shot and in-context learning abilities across diverse generation tasks. It is particularly adept at cross-domain knowledge integration and producing coherent long-form text.
- **Claude-Sonnet-4.5**: A proprietary model from Anthropic, optimized for precise instruction following and long-context reasoning. It is well-suited for document-level text processing, summarization, and complex conversational tasks, balancing capability with safety and reliability.
- **Gemini-3-Pro-Preview**: Google's advanced proprietary language model, demonstrating competitive performance in reasoning, multimodal understanding, and long-context generation. It shows particular strength in structured output generation and cross-lingual tasks.

Inference Prompt with ICL and CoT

Rewrite the following complex long text given by User into a simpler version in order to make it easier to understand by non-native speakers of English. You can do so by replacing complex words with simpler synonyms (i.e. paraphrasing), deleting unimportant information (i.e. compression), and/or splitting a long complex sentence into several simpler ones. The final simplified text needs to be grammatical, fluent, and retain the main ideas of its original counterpart without altering its meaning. You are supposed to give the detailed reasoning process step by step and then provide the final simplification.
Examples:
Original: {example_input}
Simplified: {example_simplification}

Now, simplify the following text:
User:{complex_long_text}
Reasoning Process:
Simplification:

Inference Prompt with ICL

Rewrite the following complex long text given by User into a simpler version in order to make it easier to understand by non-native speakers of English. You can do so by replacing complex words with simpler synonyms (i.e. paraphrasing), deleting unimportant information (i.e. compression), and/or splitting a long complex sentence into several simpler ones. The final simplified text needs to be grammatical, fluent, and retain the main ideas of its original counterpart without altering its meaning. You should only provide the final simplification.
Examples:
Original: {example_input}
Simplified: {example_simplification}

Now, simplify the following text:
User:{complex_long_text}
Simplification:

Inference Setup. For open-source LLMs, including Qwen2.5-72B-Instruct and Llama3.3-70B-Instruct, all inference is conducted using the vLLM framework (Kwon et al., 2023). For proprietary models, including DeepSeek-V3.2, Qwen3-Max, GPT-5.2, Claude-Sonnet-4.5 and Gemini-3-Pro-Preview, inference is performed through their official APIs, respectively. To construct stronger baselines, all models are evaluated under an ICL+CoT or ICL prompting (Wei et al., 2022; Dong et al., 2024) setting using the same simplification prompt to ensure fair comparison. During inference, we set the temperature to 0.5. For traditional text simplification methods, we follow the default parameter settings recommended in their original implementations.

C Evaluation Metrics

To provide a more comprehensive assessment of simplification quality, we adopt five complementary metrics, including SARI, D-SARI, FKGL, BERTScore, and LENS. These metrics jointly evaluate different aspects of simplification, covering lexical and structural edit operations, semantic preservation, overall quality and readability.

- **SARI**: SARI is a holistic metric for simplification quality, focusing on lexical edit operations. It measures the F1 scores of n-grams that are added, retained, or deleted when comparing the simplification with the source and reference text. SARI explicitly rewards appropriate edits and penalizes unnecessary copying, making it well suited for evaluating structural simplification quality (Xu et al., 2016).
- **D-SARI**: D-SARI extends the original SARI metric to better handle document-level simplification by penalizing its three components. Unlike SARI, D-SARI evaluates n-gram edits at the document scope, enabling more accurate assessment of global simplification behaviors such as content deletion, reorganization, and redundancy reduction (Sun et al., 2021).
- **FKGL**: The Flesch–Kincaid Grade Level (FKGL) measures the readability of the input text. It is computed based on sentence length and word syllable counts and lower scores indicate simpler and more readable simplification (Kincaid et al., 1975).
- **BERTScore**: BERTScore evaluates semantic similarity by computing token-level cosine similarity in a contextual embedding space. Given contextual embeddings from a pretrained encoder, BERTScore computes precision, recall, and F1 by greedily matching tokens between the hypothesis and reference text (Zhang et al., 2020).
- **LENS**: LENS (**L**earnable **E**valuation **M**etric for **T**ext **S**implification) is a reference-based evaluation metric specifically designed for text simplification. It leverages a model trained on complex–simple pairs with human quality annotations to produce a score ranging from 0 to 100, which predicts human judgments of simplification quality by jointly considering meaning preservation, grammaticality and simplicity (Maddela et al., 2023).

Domain	Train	Validation	Level
Medical	3,568	411	Paragraph
Wikipedia	3,600	400	Paragraph

Table 8: Training and validation data statistics used for paragraph-level training of Simplify-Pro.

D Implementation Details

In both Medical and Wikipedia domains, only a few thousand parallel paragraph samples are used for training, as summarized in Table 8. This setup allows us to assess whether Simplify-Pro can achieve strong simplification performance without relying on large-scale annotated data, thereby highlighting the efficiency of the proposed framework. More-

over, a reasoning-oriented prompt is designed to provide a more explicit task description and encourage the model to produce a detailed, step-by-step reasoning process before generating the final simplification. Our implementation is based on the verl framework (Sheng et al., 2025). We set the KL penalty coefficient β to 0 and raise clip-ratio-high to 0.28. These settings follow observations from DAPO, which indicate that KL regularization may overly constrain exploration in LLMs, while a larger clip-ratio-high enables stronger weighting of low-probability tokens during optimization (Yu et al., 2025). Training is conducted with a batch size of 16, sampling temperature of 1.0, constant learning rate of 1e-6, and maximum generation tokens of 2,048. All models are trained for 2 epochs on 8 NVIDIA A800 80G GPUs. For SFT training, we use the LLaMA-Factory framework with the following configuration: Lora Rank = 128, Lora Alpha = 256, Lora Target = all, epochs = 3, learning rate = 1e-4, and batch size = 16.

Instruction-guided Prompt for Training

You are an Expert in long text simplification. Your task is to rewrite the complex long text given by User into a simpler version in order to make it easier to understand by non-native speakers of English. The final simplified text needs to be grammatical, fluent, and retain the main ideas of its original counterpart without altering its meaning. You should give the detailed reasoning process step by step and then provide the final answer. The detailed reasoning process and final answer must be strictly enclosed within <think></think>and <answer></answer>tags, respectively, i.e., <think>reasoning process here </think><answer>final answer here </answer>.
User:{complex_long_text}
Expert:

In our framework, we use K = 4 candidate generations per input. For the selective refinement, a structured, stage-specific prompt is designed to support multi-candidate selection and refinement. The prompt integrates task instructions, quality criteria, and intermediate constraints tailored to each stage. Due to the increased prompt complexity and the need for stable long-context reasoning, all inference is performed using GPT-4o (Achiam et al., 2023) with the temperature set to 0.3.

Prompt for Selective Refinement

You are a professional lexical simplification editor. Your task is to select the most suitable candidate and perform proper lexical simplification for complex vocabularies.
For each simplification candidate, corresponding metric-assisted analysis (MA) is provided below the candidate to help you better understand its structural changes and content transformations. Structural changes are assessed through the proportions of additions, deletions and copies. And content transformations are analyzed using a combination of Levenshtein Similarity and Lexical Complexity. You may use these values as supportive signals, but your decision should still primarily follow correctness, safety and the lexical-simplification principle.
The proportions of additions, deletions and copies reflect how the simplified text differs from the original input in terms of editing operations.
Levenshtein Similarity measures the extent of editing required to transform the original text into the simplification.
Lexical Complexity reveals whether simplification leads to the use of simpler and more accessible vocabulary.

Candidates Set:
{paired simplification and MA}

Strict Rules:
1. Maintain overall structure and avoid sentence merging/splitting/reordering/rewriting as much as possible.
2. No adding or removing facts/entities/experimental data.
3. Focus on performing lexical simplification, including but not limited to replacing the uncommon/complex/awkward word with a simpler exact-meaning synonym and fixing clear grammar or word errors.

Output Requirement:
Return only the final simplification without explanation or labels.

Algorithm 1: Multi-stage Pipeline of Simplify-Pro

Input: Complex long text X

Output: Final simplified text Y^*

Stage 1: Paragraph-level Training

Initialize paragraph-level policy π_θ from a pretrained LLM.

for *each training step* **do**

Sample a mini-batch $\{(q_b, a_b)\}_{b=1}^{B} \sim D$.

for $b = 1$ **to** B **do**

Sample a group of responses $\{o_{b,i}\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot \mid q_b)$.

for $i = 1$ **to** G **do**

Compute total reward $R_{b,i}$.

for $i = 1$ **to** G **do**

$$\hat{A}_{b,i,t} = \frac{R_{b,i} - \text{mean}\Big(\{R_{b,j}\}_{j=1}^{G}\Big)}{\text{std}\Big(\{R_{b,j}\}_{j=1}^{G}\Big)}$$

$$J_{\text{GRPO}}(\theta) = \text{E}_{(q_b,a_b)\sim D,\ \{o_{b,i}\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q_b)}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_{b,i}|}\sum_{t=1}^{|o_{b,i}|}\min\Big(r_{b,i,t}(\theta)\hat{A}_{b,i,t}, \text{clip}(r_{b,i,t}(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_{b,i,t}\Big) - \beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})\Bigg]$$

Stage 2: Simplification Generation

for $i = 1$ **to** M **do**

$V_i = Encoder(S_i)$;

for $i = 1$ **to** $M - 1$ **do**

$$Sim_i = \frac{V_i^\top V_{i+1}}{\|V_i\|\,\|V_{i+1}\|};$$

Identify initial boundaries where $Sim_i < \tau$;

Apply greedy merging with soft length constraints to obtain blocks $\{B_j\}_{j=1}^{N}$;

for $j = 1$ **to** N **do**

$\hat{B}_j \sim \pi_\theta(\cdot \mid B_j)$;

Concatenate $\{\hat{B}_j\}_{j=1}^{N}$ to form K candidates $\{Y_i\}_{i=1}^{K}$;

Stage 3: Metric-assisted Analysis

for $i = 1$ **to** K **do**

$$(P_{edit}^{(i)})_{edit\in\{add,del,copy\}} = \frac{N_{edit}^{(i)}}{\sum_{edit\in\{add,del,copy\}} N_{edit}^{(i)}};$$

$$(R_{compress}^{(i)}, LevSim^{(i)}, LexC^{(i)}) = \Big(1 - \frac{|Y_i|}{|X|},\ 1 - \frac{d_{\text{lev}}(X, Y_i)}{\max(|X|, |Y_i|)},\ -\frac{1}{|Y_i|}\sum_{w\in Y_i}\log(f(w) + \delta)\Big);$$

$$A(Y_i) = \{P_{edit}^{(i)}, R_{compress}^{(i)}, LevSim^{(i)}, LexC^{(i)}\};$$

Stage 4: Selective Refinement

Construct paired inputs $\{(Y_i, A(Y_i))\}_{i=1}^{K}$.

Prompt a large language model to select the best candidate Y^*.

Perform lexical-level refinement on Y^* under strict structural and factual constraints.

return Y^*

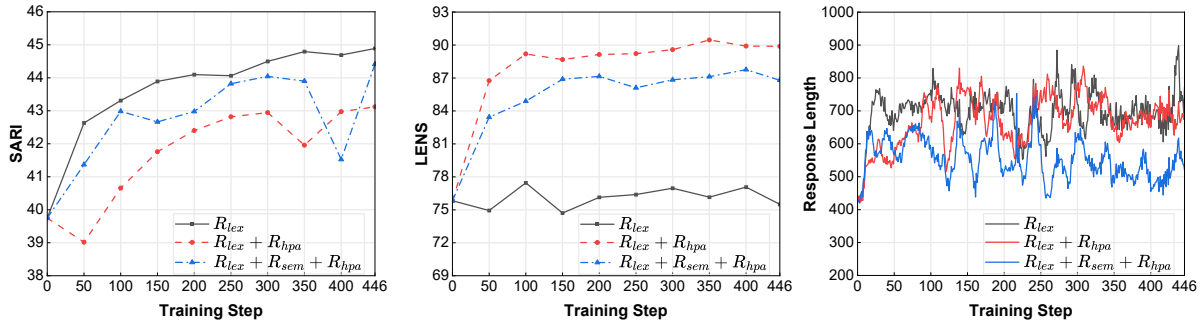

Figure 4: Paragraph-level Training dynamics of simplification models in the Medical domain (using R_{lex}, $R_{lex} + R_{hpa}$ and $R_{lex} + R_{sem} + R_{hpa}$, respectively). Here we show SARI and LENS progression on validation sets, and average response length changes over training steps.

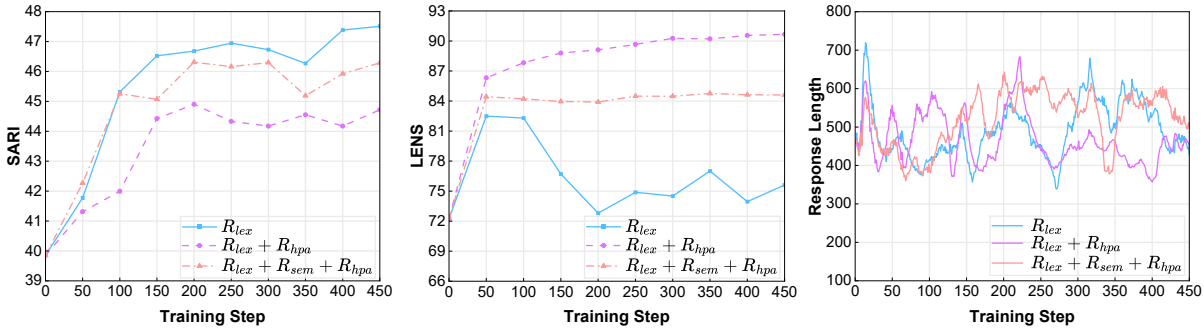

Figure 5: Paragraph-level Training dynamics of simplification models in the Wikipedia domain (using R_{lex}, $R_{lex} + R_{hpa}$ and $R_{lex} + R_{sem} + R_{hpa}$, respectively). Here we show SARI and LENS progression on validation sets, and average response length changes over training steps.

E Overall Algorithm for Simplify-Pro

Algorithm 1 presents the complete implementation pipeline of Simplify-Pro, integrating paragraph-level training, simplification generation, metric-assisted analysis, and selective refinement into a multi-stage framework.

Following GRPO-style optimization, the policy update is performed at the token level. For the i-th sampled output sequence $o_i = (o_{i,1}, \ldots, o_{i,|o_i|})$, the token-level importance ratio at step t can be calculated as

$$r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t} \mid x, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid x, o_{i,<t})}, \tag{12}$$

where $\pi_{\theta_{\text{old}}}$ denotes the behavior policy used to sample the candidate sequences.

Specifically, the advantage and KL divergence approximation term is computed as

$$\hat{A}_{i,t} = \frac{R_i - \text{mean}\left(\{R_j\}_{j=1}^{G}\right)}{\text{std}\left(\{R_j\}_{j=1}^{G}\right)} \tag{13}$$

$$D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}}) = \frac{\pi_{\text{ref}}(o_i \mid q)}{\pi_\theta(o_i \mid q)} - \log \frac{\pi_{\text{ref}}(o_i \mid q)}{\pi_\theta(o_i \mid q)} - 1 \tag{14}$$

F Discussion of Paragraph-level Training

Figure 4 illustrates the training dynamics of paragraph-level simplification models in the Medical domain under three reward configurations (R_{lex}, $R_{lex} + R_{hpa}$, and $R_{lex} + R_{sem} + R_{hpa}$). We report the progression of SARI and LENS on validation sets, together with the evolution of average response length across training steps.

From the SARI curve, optimizing with R_{lex} leads to the most stable and monotonic improvement, reaching and maintaining the highest score among all settings after the early stage of training. This indicates that a pure lexical and structural reward signal strongly pushes the model toward edit operations favored by SARI (addition/deletion/retainment). However, this gain fails to bring about better human preference alignment, where the corresponding LENS scores remain relatively lower and fluctuate around a relatively flat band.

In contrast, incorporating R_{hpa} leads to a pronounced early increase in LENS, which remains at a consistently high level throughout subsequent training. This behavior demonstrates that introducing an explicit quality-oriented metric can effectively suppress overly aggressive edits. However, this improvement is accompanied by a slower SARI progression. Under the $R_{lex} + R_{hpa}$ setting, SARI

increases more gradually and remains below the R_{lex} throughout most of training, reflecting a conservative editing strategy that balances fewer structural edits against human preference alignment.

When all three components are combined ($R_{lex}+R_{sem}+R_{hpa}$), the model exhibits a more balanced training trend. SARI improves steadily and reaches competitive values in the mid-to-late stage, while LENS remains substantially higher than the R_{lex} setting. Such pattern indicates that the triple mixed reward provides a more comprehensive learning signal spanning lexical edit operations, semantic similarity, and human preference alignment. We also observe a brief late-stage instability in SARI for both mixed reward settings (a transient drop near the end of training), which indicates the model policy temporarily shifts toward outputting simplifications that preserve overall quality at the cost of better lexical edit operations.

The response length trajectories further reveal how different reward signals shape generation behavior. Under R_{lex}, the average response length quickly rises from the initial short outputs to a relatively long band and keeps oscillating with large spikes, suggesting that the model often produces longer simplification and continues to experiment with more detailed reasoning steps. Adding R_{hpa} noticeably moderates this behavior. The length curve becomes less extreme and tends to stabilize around a balanced range, which supports the view that human preference alignment reward discourages unnecessary edits and reduces the tendency to introduce extra content. Most notably, the triple mixed reward ($R_{lex}+R_{sem}+R_{hpa}$) yields the shortest and most controlled outputs overall, with the length curve settling into a lower band and exhibiting smaller oscillations in the later training stage. This pattern suggests that incorporating semantic and human preference supervision encourages the model to converge to more concise reasoning or simplification, whereas SARI-only optimization may implicitly reward longer paraphrastic rewrites with more edit operations which may fail to contribute to better simplification quality.

Figure 5 reports the paragraph-level training dynamics in the Wikipedia domain under the same three reward configurations. Overall, several trends observed in the Medical domain are also present here. In particular, optimizing with R_{lex} yields the fastest and highest improvement in SARI, while exhibiting weaker LENS performance. Meanwhile, incorporating R_{hpa} or the triple mixed reward leads to substantially better LENS scores. The similar patterns suggest that the fundamental trade-offs induced by different reward designs are stable and not domain-specific.

Despite these similarities, paragraph-level training in the Wikipedia domain exhibits several distinctive behaviors compared to the Medical setting. First, the final SARI scores achieved are relatively higher across all reward configurations. This indicates that Wikipedia simplification allows the model to reach a higher level of structural edit quality under the same training setups. Also, Wikipedia text is well suited to edit-level simplifications, as its more accessible content and fewer technical terms allow SARI metric to be optimized more effectively.

Second, response length dynamics differ substantially across the two domains. In the Medical domain, simplification outputs tend to be longer and show greater variability, reflecting the need for more involved reasoning and careful reformulation when handling specialized content. By contrast, Wikipedia simplification generally results in shorter outputs that stabilize earlier during training, as its content is typically less technical and allows for more direct paraphrasing without complex domain-specific reasoning.

These observations highlight that paragraph-level training dynamics are shaped not only by reward design but also by domain characteristics. Medical simplification is more challenging in general, as it involves more experimental explanations and specialized content. In contrast, Wikipedia simplification follows more regular textual patterns and allows more flexible rewriting. These differences motivate mixed reward designs that adapt to domain-specific needs while preserving training stability.

G Supplementary Experimental Results

Performance Across Models. Table 9 further reports the performance comparison on paragraph-level simplification in the Medical domain, where the first training stage of Simplify-Pro is instantiated with two different LLMs, Qwen2.5-14B-Instruct and Llama3.1-8B-Instruct. Notably, the trends observed at the document level from Table 4 remain consistent at the paragraph level. Across all three framework variants, Simplify-Pro with Qwen2.5-14B-Instruct still achieves higher scores on most metrics. Specifically, the former yields

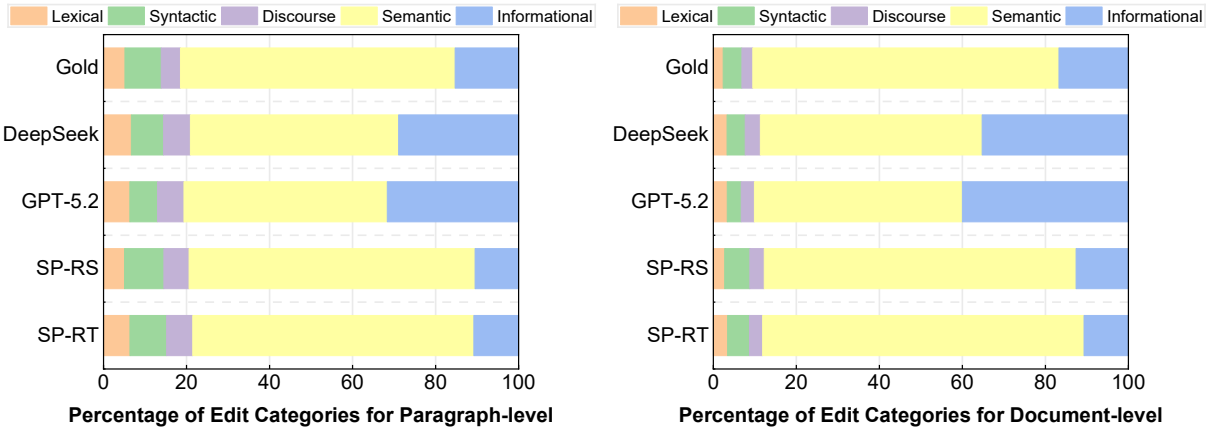

Figure 6: Distribution of five edit categories for paragraph- and document-level simplifications in the Wikipedia domain. Gold, SP-RS and SP-RT denote reference simplifications, Simplify-Pro-RS and Simplify-Pro-RT, respectively.

Method	SARI	D-SARI	FKGL	BERT	LENS
Qwen2.5-14B-Instruct for paragraph-level training					
Simplify-Pro-RS	44.99	32.89	9.99	25.47	77.15
Simplify-Pro-RD	43.42	30.00	6.29	21.36	90.14
Simplify-Pro-RT	44.64	36.19	8.74	28.49	87.25
Llama3.1-8B-Instruct for paragraph-level training					
Simplify-Pro-RS	45.24	32.28	11.05	24.81	74.47
Simplify-Pro-RD	42.83	29.58	5.91	17.20	89.14
Simplify-Pro-RT	43.76	35.66	8.74	27.43	85.50

Table 9: Performance comparison on paragraph-level simplification in the Medical domain. For our framework, Qwen and Llama are used for the paragraph-level training stage, respectively.

Method	SARI	D-SARI	FKGL	BERT	LENS
Paragraph-level Simplification					
Simplify-Pro-RT	**47.59**	**37.23**	3.28	**34.64**	84.56
w/o MA	46.77	36.79	**3.08**	33.80	**84.77**
w/o MA&SR	46.90	36.68	3.56	33.66	83.86
Document-level Simplification					
Simplify-Pro-RT	**45.83**	37.57	**3.38**	28.61	**83.80**
w/o MA	45.42	**37.67**	3.87	**28.82**	83.18
w/o MA&SR	45.28	37.09	3.59	27.90	82.88

Table 10: Ablation study on Simplify-Pro-RT's simplification performance in the Wikipedia domain. MA and SR refer to the stage of quality analysis and selective refinement, respectively. The best results are in bold.

greater gains in BERTScore and LENS compared to the latter (up to +4.16/+2.68), which suggests that the larger model capacity is beneficial for semantic preservation while maintaining human preference alignment in paragraph-level simplification. However, the performance gap remains relatively small, and Simplify-Pro with Llama3.1-8B-Instruct occasionally attains competitive or even slightly better scores on individual metrics such as SARI or FKGL. Overall, the results above complement the document-level analysis in Section 4.3, providing further evidence that our proposed framework generalizes well across different models.

Ablation Study. Table 10 presents the ablation study of Simplify-Pro-RT in the Wikipedia domain, reporting results at both paragraph and document levels. At the paragraph level, removing the metric-assisted analysis (MA) stage leads to a clear degradation on core structural and semantic metrics, where SARI drops from 47.59 to 46.77 (-0.82), D-SARI decreases from 37.23 to 36.79 (-0.44), and BERTScore declines from 34.64 to 33.80 (-0.84). This pattern indicates that MA provides useful guidance for candidate comparison and helps retain structural edit quality while preserving semantic similarity. Further removing selective refinement (SR) causes an additional drop in human preference alignment, with LENS decreasing from 84.56 to 83.86 (-0.70), while SARI and D-SARI remain below the complete stage. This suggests that SR plays an important role in improving human preference alignment. At the document level, the full framework achieves the best SARI (45.83), while removing MA reduces SARI to 45.42 (-0.41) and further removing SR reduces it to 45.28 (-0.55). At the same time, LENS declines from 83.80 to 83.18 (-0.62) when MA is removed and further to 82.88 (-0.92) when both MA and SR are removed. Although the ablated variants occasionally obtain slightly better scores on isolated metrics (e.g., D-SARI or BERTScore), such gains are at the cost of other key dimensions, resulting in less balanced simplification behavior. In generation, these ablation results confirm that both MA and SR are essential stages of Simplify-Pro-RT, jointly contributing to more balanced and effective long text simplification.

Simplification Analysis. Figure 6 shows the dis-

tribution of five edit categories for paragraph- and document-level simplifications in the Wikipedia domain. Similar to observations in the Medical domain, Simplify-Pro variants (SP-RS and SP-RT) produce edit distributions that are much closer to the gold human-annotated references than strong proprietary LLMs such as DeepSeek-V3.2 and GPT-5.2. Meanwhile, baseline LLMs display different editing behaviors. They tend to overemphasize lexical or informational edits while underperforming on semantic-level operations, leading to distributions that deviate substantially from the gold reference. The discrepancies indicate that, although strong proprietary LLMs can generate fluent outputs, they often fail to adopt the balanced and human-like simplification strategies. These results demonstrate that Simplify-Pro captures the appropriate balance between several edit operations and produces human-aligned simplification strategies across domains, supporting its effectiveness and robustness for long text simplification.

H Edit Definitions

Five edit categories used for simplification analysis can be described as follows:

- Lexical edits focus on simplifying word-level units by replacing rare or technical terms, including single words or short phrases, with simpler and more familiar alternatives (Paetzold, 2015; Paetzold and Specia, 2016).
- Syntactic edits aim to reduce structural complexity through operations such as sentence splitting, sentence fusion, syntactic deletion, reordering clauses, or adjusting sentence form (Zhu et al., 2010; Lu et al., 2021).
- Discourse edits are dedicated to improving text coherence at the multi-sentence level by modifying how information is connected and organized across sentences. Typical operations include clarifying sentence relationships, reordering content to present foundational information before more advanced concepts and replacing repeated or implicit expressions (Devaraj et al., 2022; Agrawal and Carpuat, 2022).
- Semantic edits concentrate on modifying the content of a text by adding or removing information to improve overall readability at the document level. Such edits include removing details that are not essential for basic comprehension, introducing additional background information or illustrative examples that help a broader audience grasp the main ideas of the content, and substituting low-level details in exchange for higher-level descriptions (Alva-Manchego et al., 2019b; Yanamoto et al., 2022).
- Informational edits cover modifications that do not directly contribute to text simplification, but instead address auxiliary changes related to content correctness or formatting. Examples include correcting factual inaccuracies, cleaning noisy or irregular text, adjusting formatting conventions and inserting secondary information that is not essential for simplified understanding (Alva-Manchego et al., 2019a).

To support the analysis of simplification behaviors, we employ a joint grouping-and-categorization model, namely BIC (BI-Category), which has been specifically trained for edit operation identification (Laban et al., 2023). Given the original input and corresponding simplified text, it predicts both edit boundaries and edit categories simultaneously, enabling accurate identification of complex and interleaved edit operations that commonly occur in long text simplification. In our experiments, a total of 19 fine-grained edit types are identified automatically using the BIC model. For clearer presentation and more convenient analysis, all fine-grained operations are finally grouped into the five high-level simplification categories described above.

I Discussion of Simplification Generation

Figure 7 and Figure 9 present qualitative comparisons of CoT steps during simplification generation. Across both domains, Simplify-Pro-RS follows a more structured reasoning process, explicitly identifying and retaining key factual elements from the complex input before reformulating them in simpler descriptions. In contrast, Simplify-Pro-RT adopts a more reader-oriented reasoning pattern, selectively prioritizing high-level conclusions and intuitive explanations while omitting fine-grained details. The distinction between reasoning strategies becomes more evident across domains. In the Medical domain, RS reasoning tends to track uncertainty-related information and evidence strength, reflecting a cautious and detailed interpretation of complex medical content. Meanwhile, RT reasoning places greater emphasis on identifying concise, patient-facing explanations, and deleting technical details. In the Wikipedia domain, texts are relatively accessible with less specified terminology, and RT reasoning more fre-

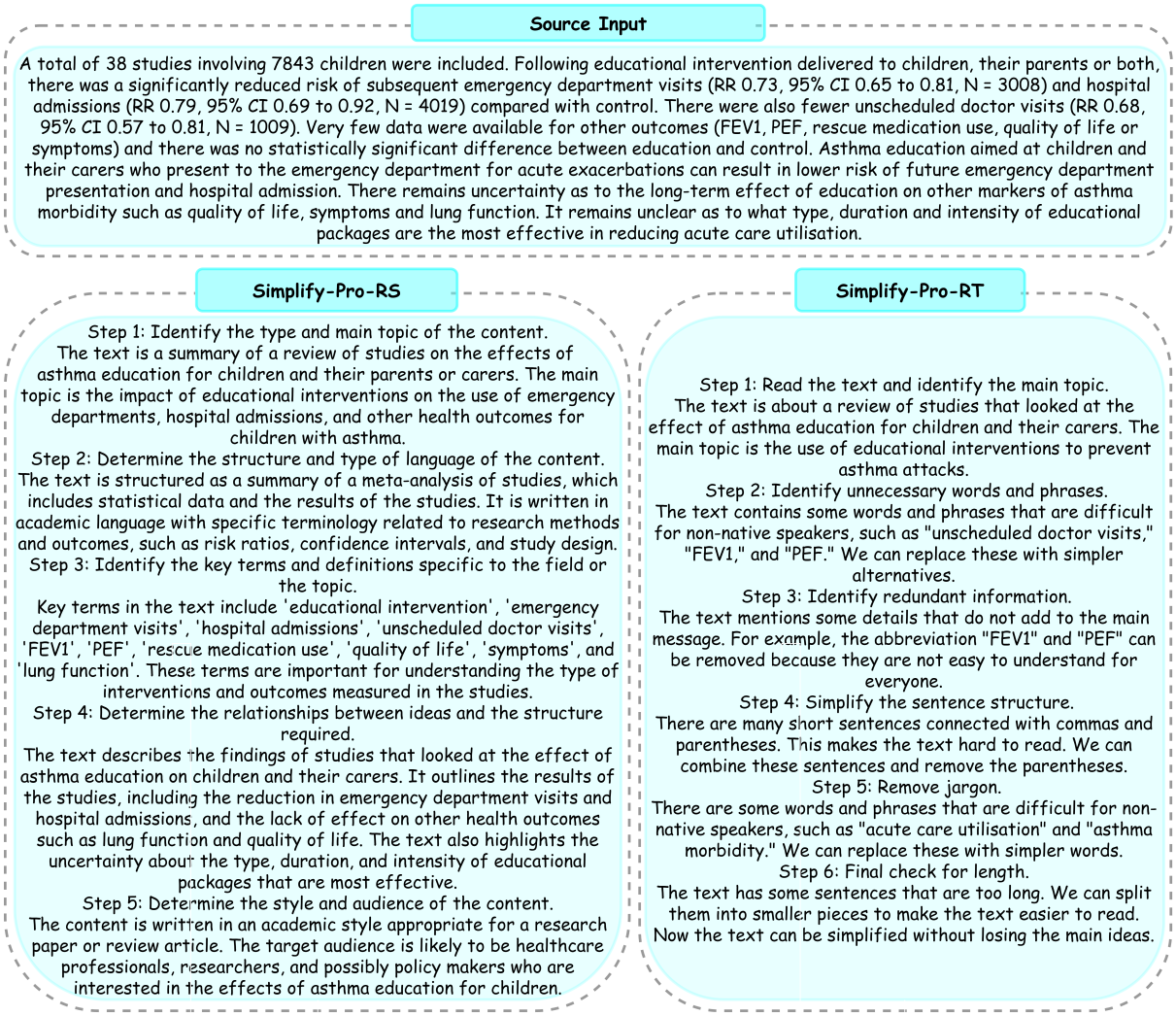

Figure 7: Comparison on CoT in the Medical domain during the stage of simplification generation. Here we show the detailed reasoning steps from Simplify-Pro-RS and Simplify-Pro-RT.

quently compresses extensive background information into core conceptual descriptions. Moreover, RS continues to reason over a broader range of contextual elements. Together, these observations reflect how reasoning strategies adapt to different reward signals and domain characteristics.

J Case Study

Figure 8 and Figure 10 present representative paragraph-level case studies in the Medical and Wikipedia domains, showing the original complex input, the human-annotated reference, and simplifications generated by Simplify-Pro-RS and Simplify-Pro-RT. Within the Medical domain, RS outputs tend to retain relevant details that appear in the reference, including contextual qualifiers and conditions that delimit the applicability of the findings. On the other hand, RT outputs show a different simplification orientation by consolidating the content into a concise set of high-level medical statements. This abstraction behavior highlights the strength in distilling complex medical content into clear, accessible simplification text. A similar pattern can also be observed within the Wikipedia domain. When compared to the reference, RS-generated outputs exhibit differences from the reference primarily at the level of linguistic formulation. The outputs often reorganize or rephrase informational content while preserving a comparable set of descriptive elements, which is reflected in wording choices and sentence structure. By comparison, RT-generated outputs place greater focus on information selection, emphasizing a smaller set of descriptive facets.

Besides, the case studies also reveal clear differences between Medical and Wikipedia simplification. In the Medical domain, simplified outputs tend to vary with respect to how much contextual and specialized information is retained, which re-

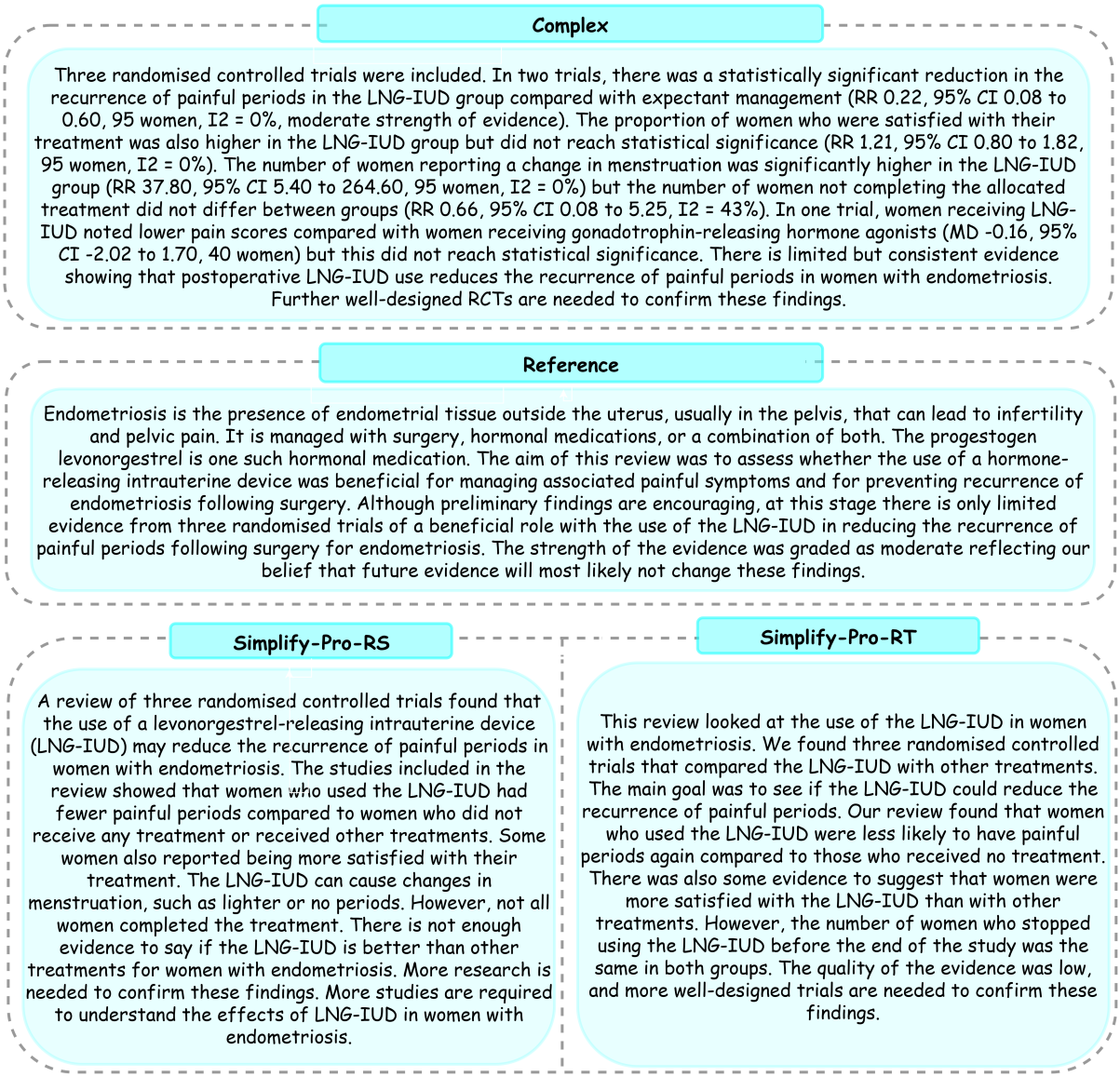

Figure 8: Case study of paragraph-level simplification results generated by our framework in the Medical domain. For each example, we show the original complex input with corresponding reference simplification, alongside the simplified outputs produced by Simplify-Pro-RS and Simplify-Pro-RT.

flects the need to balance explanation of experimental facts with better readability. In contrast, Wikipedia simplifications more often differ from reorganization of descriptive content, while preserving the mentioned concepts. As a result, Medical simplification is characterized by adjustments in informational granularity, and Wikipedia simplification primarily involves changes in textual structure and presentation.

K More Simplification Examples

We further provide additional document-level simplification cases generated by Simplify-Pro. These examples are presented after Figure 10 to illustrate our framework's performance on document-level inputs, complementing the paragraph-level case studies in Appendix J.

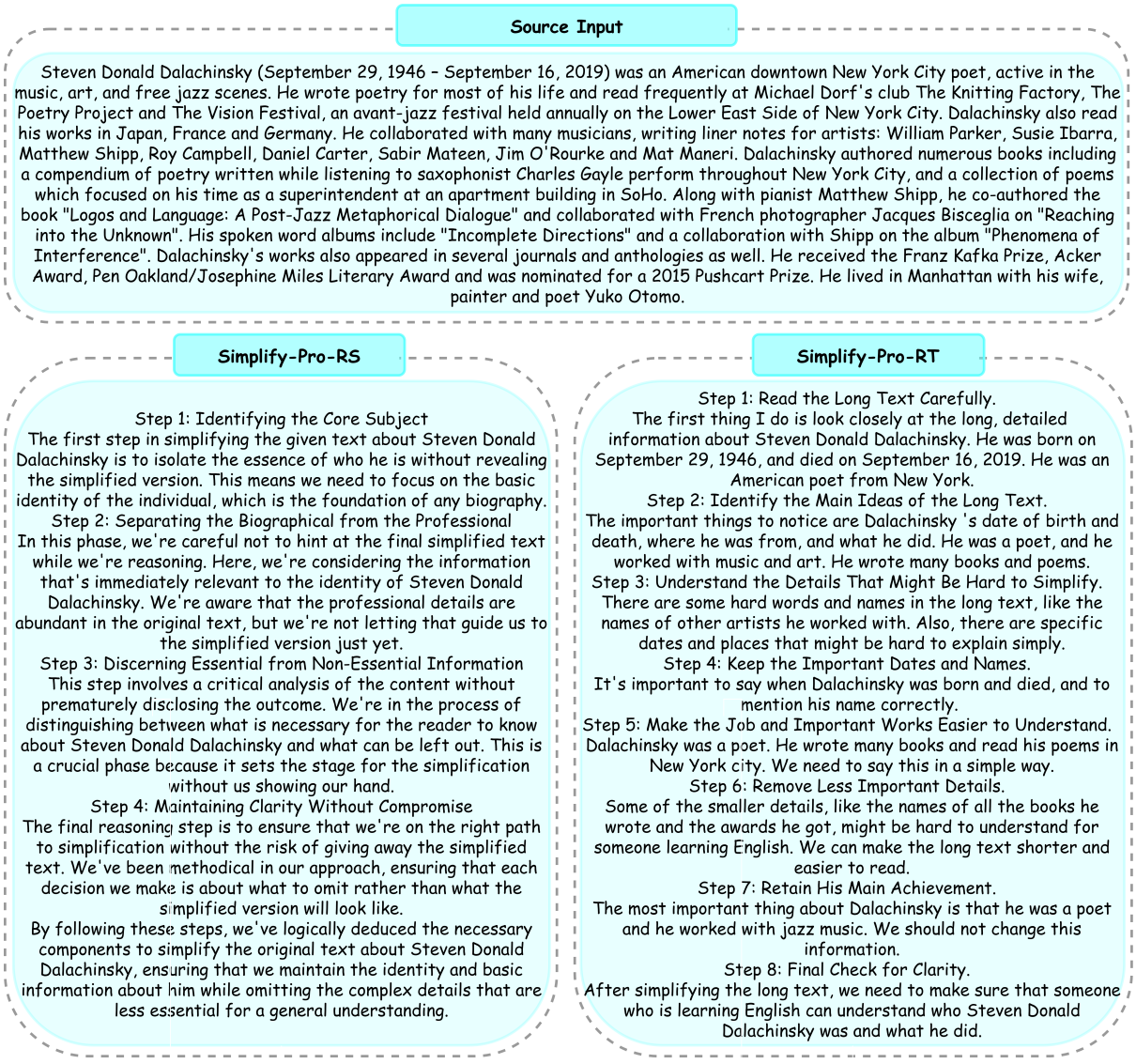

Figure 9: Comparison on CoT in the Wikipedia domain during the stage of simplification generation. Here we show the detailed reasoning steps from Simplify-Pro-RS and Simplify-Pro-RT.

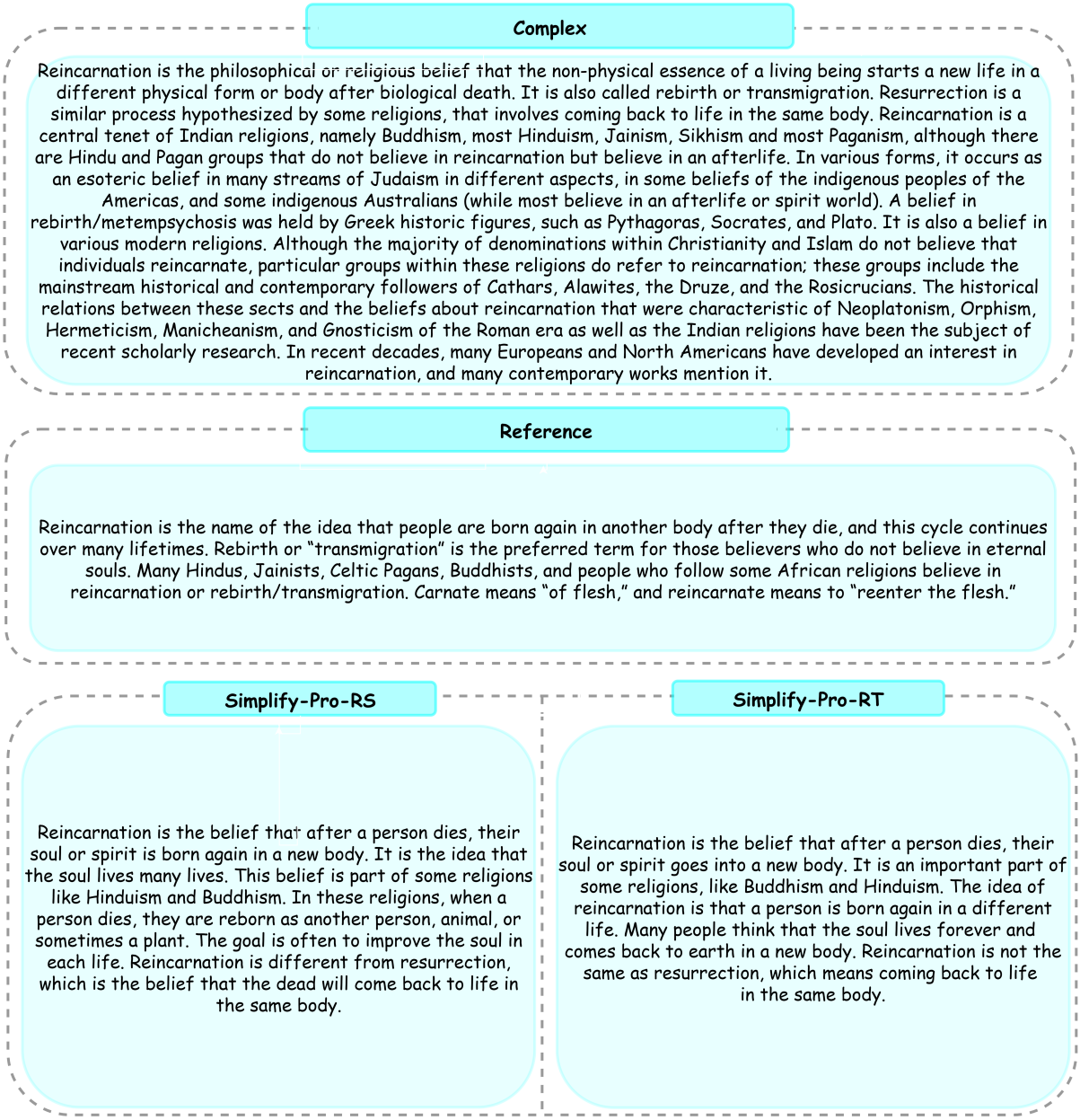

Figure 10: Case study of paragraph-level simplification results generated by our framework in the Wikipedia domain. For each example, we show the original complex input with corresponding reference simplification, alongside the simplified outputs produced by Simplify-Pro-RS and Simplify-Pro-RT.

Document-level Complex Long Text (Medical)

Diagnoses of endometrial cancer are increasing secondary to the rising prevalence of obesity. Obesity plays an important role in promoting the development of endometrial cancer, by inducing a state of unopposed oestrogen excess, insulin resistance and inflammation. It also affects treatment, increasing the risk of surgical complications and the complexity of radiotherapy planning, and may additionally impact on subsequent survival. Weight-loss interventions have been associated with improvements in breast and colorectal cancer-specific survival as well as a reduction in the risk of cardiovascular disease, a frequent cause of death in endometrial cancer survivors.

To determine the impact of weight-loss interventions, in addition to standard management of endometrial cancer, on overall survival and the frequency of adverse events, we include an assessment of weight-loss interventions on endometrial cancer-specific survival, weight loss achieved, cardiovascular event frequency and quality of life both overall and stratified according to patient body mass index (BMI), where possible. Randomised controlled trials (RCTs) of interventions to facilitate weight loss in overweight or obese women undergoing treatment for, or previously treated for, endometrial cancer were selected. Two review authors independently selected studies, assessed trial quality, and extracted data with disagreements resolved by a third review author.

We included three RCTs, randomising a total of 161 overweight and obese women with endometrial cancer. All studies compared combined behavioural and lifestyle interventions to facilitate weight loss through dietary modification and increased physical activity. The included RCTs were of low or very low quality, due to high risk of bias by failing to blind participants, personnel and outcome assessors, and significant loss to follow-up (attrition rate up to 29%). Combined behaviour and lifestyle interventions were not associated with improved overall survival (risk ratio (RR mortality), 0.23 95% confidence interval (CI) 0.01 to 4.55, P = 0.34, one RCT, 37 participants; very low-certainty evidence) compared with usual care at 24 months. There was no evidence that such interventions were associated with improvements in cancer-specific survival or cardiovascular event frequency as no cancer-related deaths, myocardial infarctions or strokes were reported in the included studies. None of the included RCTs reported data for the outcome of recurrence-free survival. Combined behaviour and lifestyle interventions were not associated with significant weight loss at either six months (mean difference (MD) -1.88 kg, 95% CI -5.98 to 2.21 kg, P = 0.37, three RCTs, 131 participants, I 2 = 0%; low-certainty evidence)or 12 months (MD -8.98 kg, 95% CI -19.88 to 1.92 kg, P = 0.11, two RCTs, 91 participants, I 2 = 0%; very low-certainty evidence) when compared with usual care. Combined behaviour and lifestyle interventions were not associated with increased quality of life, when measured using either the SF-12 Physical Health questionnaire or FACT-G at six months (FACT-G MD 2.51, 95% CI -5.61 to 10.64, P = 0.54, two RCTs, 95 participants, I 2 = 83%; very low-certainty evidence), or by FACT-G alone at 12 months (MD 2.77, 95% CI -0.65 to 6.20, P = 0.11, two RCTs, 89 participants, I 2 = 0%; very low-certainty evidence) when compared with usual care. No serious adverse events like hospitalisation were reported in included trials. Lifestyle and behavioural interventions were associated with a higher risk of musculoskeletal symptoms (RR 19.03, 95% CI 1.17, 310.52, P = 0.04, two RCTs, 91 participants; low-certainty evidence).

There is currently insufficient high-quality evidence to determine the effect of combined lifestyle and behavioural interventions on survival, quality of life, or significant weight loss in women with a history of endometrial cancer compared to those receiving usual care. The limited evidence suggests that there is little or no serious or life-threatening adverse effects due to these interventions, although musculoskeletal problems were increased, presumably due to increased activity levels. Our conclusion is based on low- and very low-quality evidence from a small number of trials and very few patients. We therefore have very little confidence in the evidence: the true effect of weight-loss interventions in obese women with endometrial cancer is currently not known. Further methodologically-rigorous, adequately-powered RCTs are required with follow-up of 5 to 10 years duration. These should focus on the effects of varying dietary modification regimens, pharmacological treatments associated with weight loss and bariatric surgery on survival, quality of life, weight loss and adverse events.

Reference Simplification

Weight-loss interventions in endometrial cancer survivors or womb cancer is a common cancer in women and the number of cases is rising. This is due, in part, to increasing levels of obesity, which is a major risk factor for the disease. Whilst survival following endometrial cancer is generally excellent if diagnosed early, affected women are more likely to die early due to an increased risk of heart attacks and strokes and to have poorer quality of life. This review assessed the evidence for weight-loss interventions in overweight and obese endometrial cancer survivors to determine whether they were of benefit compared with usual care. We included three randomised controlled trials in which women were allocated at random to receive one of several interventions (treatments) and which involved 161 obese participants. The trials were conducted in the USA and the UK. All compared lifestyle advice (diet and exercise) plus self-help techniques (to encourage adherence to the advice) with usual care. The evidence is current to January 2018. We found no benefit for endometrial cancer survivors from receiving lifestyle advice in terms of survival, cardiovascular events or quality of life, though such interventions were not associated with significant or serious harms to participants. They did, however, report higher rates of musculoskeletal symptoms, presumably due to increases in physical activity. Whilst some women lost weight with these interventions, others did not, meaning that overall there was little or no benefit. The quality of included studies was, however, low or very low and all were small in terms of the number of participants and not designed to specifically look at the effect of their intervention on survival. Additional high-quality studies are required in this field and currently there are five ongoing trials.

Simplification by Simplify-Pro

Endometrial cancer is a type of cancer that affects the lining of the womb. It is becoming more common, especially in women who are overweight or obese. Obesity can increase the risk of endometrial cancer. It can also make treatment harder and may affect how long people survive. Some studies have shown that losing weight can help prevent other health problems, such as heart disease. Researchers wanted to know if weight-loss interventions could help women with endometrial cancer. They looked at three randomised controlled trials to see if these interventions made a difference. The trials compared women who received usual care with those who were given diet and exercise programs. The goal was to find out if these programs could improve survival, quality of life, and weight loss. A total of 161 women took part in the studies. The researchers found that combined behaviour and lifestyle interventions did not make a big difference in overall survival. There was no evidence that these interventions helped reduce the risk of cancer or heart problems. Most of the women in the studies did not lose much weight. Some people had muscle and joint pain, but there were no serious side effects. The quality of the evidence was very low because the trials were small and not well done. There was a lot of uncertainty about the effects of weight-loss interventions in women with endometrial cancer. The studies were not good enough to tell us whether these programs were helpful or not. This is important because many women with endometrial cancer are overweight. We need more research to understand how weight loss can affect people with this type of cancer. The trials should look at different ways to help people lose weight, such as changing their diet or using medicine. This information is needed to make better decisions about treatment.

Document-level Complex Long Text (News)

Vienna is the worlds' best city to live in, Baghdad is the worst and London, Paris and New York do not even make it into the top 35, according to international research into quality of life. German-speaking cities dominate the rankings in the 18th Mercer Quality of Life study, with Vienna joined by Zurich, Munich, Dusseldorf and Frankfurt in the top seven. Paris has tumbled down the league, falling ten places to 37th, just ahead of London at 39th, almost entirely because of the citys vulnerability to terrorist attacks. The study examined social and economic conditions, health, education, housing and the environment, and is used by big companies to assess where they should locate and how much they should pay staff. Viennese-born Helena Hartlauer, 32, said she was not surprised at her citys top position. The municipalitys social democratic government has a long tradition of investing in high-quality social housing, making Vienna almost uniquely affordable among major cities.

US cities perform relatively poorly in the study, largely because of issues around personal safety and crime. The highest ranking city in the US is San Francisco, at 28th; Boston is 34th. Canadian cities, led by Vancouver, far outrank their US rivals in the table. You dont realize how safe Vienna is until you head abroad, said Hartlauer. We also have terric public transport, with the underground working 24 hours at weekends, and it only costs 1 per trip. Vienna benefited enormously from the fall of the Berlin Wall, becoming the gateway to Eastern European countries that often have historic ties to the former Austro-Hungarian empire. Our big USP is our geographical location, said Martin Eichtinger, Austrian ambassador to London, who lived in Vienna for 20 years. The fall of the Berlin Wall helped define Vienna as the hub for companies wanting to do business in Central Europe.

According to the World Bank, Austria has one of the highest figures for GDP per head in the world, just behind the US and ahead of Germany and Britain, although quite some way below neighbouring Switzerland. Zurich in Switzerland is named by Mercer as having the worlds' second highest quality of life but the Viennese say their city is far more fun. There are more students in Vienna than any other German-speaking city, said Hartlauer. Its a very fast growing, young and lively city, she added though she conceded she works for the citys tourist board. Vienna has long been overlooked by British weekend city break tourists, who instead flock to Barcelona or Berlin and tend to think of Austria as somewhere for skiing, lakes and mountains. But, after an increase in budget flights from regional British cities such as Manchester and Edinburgh, Vienna is fast catching up as a popular destination. In 2015, there were 588,000 British visitors to Vienna, up 18% on the year before. Eichtinger said London has become the number one city destination for Austrian visitors. Vienna has ranked top in the last seven published rankings, said Mercer. It scores highly in a number of categories; it provides a safe and stable environment to live in, a high level of public utilities and transport facilities and good recreational facilities. The European migrant crisis, which has seen large numbers of refugees and asylum seekers pass through Vienna en route to Germany, has had little impact on the city of nearly 1.8 million people, said Eichtinger.

London has never been in the quality-of-life top ten, says Mercer, damaged by its poor scores for air pollution, traffic congestion and climate. After London, Edinburgh is the next-ranking British city, in 46th place. Paris has suffered the biggest fall in the most recent rankings. Paris remained stable for several years but has, this year, dropped ten places in the overall ranking, said Mercer. The drop was essentially due to the terrorist attacks in 2015. However, it is important to highlight that safety issues are a very highly weighted factor within the basket so any small adjustments can have a big impact on the ranking. Auckland in New Zealand was the highest ranking English-speaking city in the survey, in third place, followed by Vancouver in fth. Australian cities also perform very highly in the survey, with Sydney 10th and Melbourne 15th. The Economist has consistently ranked Melbourne as the worlds' most liveable city, although its survey has been criticized as too Anglocentric. War and political unrest are behind all the worst-ranked cities in the world. Surprisingly, Damascus is named as only the seventh worst, ranked better than not just Baghdad but also Bangui in Central African Republic, Sanaa in Yemen, Port-au-Prince in Haiti, Khartoum in Sudan and NDjamena in Chad.

Reference Simplification

Vienna is the worlds' best city to live in, Baghdad is the worst and London, Paris and New York are not in the top 35, says an international study on quality of life. German-speaking cities do well in the 18th Mercer Quality of Life study, with Vienna, Zurich, Munich, Dusseldorf and Frankfurt in the top seven. Paris fell ten places to 37th. This was mostly because of the terrorist attacks on the city. Paris was just above London in 39th place. The study looked at the economy, health, education, housing and the environment. Big companies use the results of the study to decide where they should open of ces and factories and how much they should pay their employees. Helena Hartlauer, 32, is from Vienna. She said she was not surprised about her citys top position. For many years, Viennas government has spent money on good social housing. This makes Vienna a cheap place to live compared to other big cities. I live in a 100 square-metre apartment in a good area about 20 minutes walk from the city centre. But my rent is just 800 (625) a month. A similar apartment in London costs over 2,000 and even more in New York, which came 44th in the study. US cities do badly in the study, mostly because of worries about personal safety and crime. The US city in top position is San Francisco, in 28th position; Boston is 34th. You dont realize how safe Vienna is until you go abroad, said Hartlauer. We also have terri c public transport the underground trains run 24 hours at weekends and it only costs 1 per trip. Viennas location is very special, said Martin Eichtinger, Austrian ambassador to London, who lived in Vienna for 20 years. The fall of the Berlin Wall helped make Vienna a centre for companies who want to do business in Central Europe. Mercer says Zurich in Switzerland has the worlds second highest quality of life but the Viennese say their city is far more fun. There are more students in Vienna than any other German-speaking city, said Hartlauer. Its a very young and lively city, she added.

Simplification by Simplify-Pro

According to a study by Mercer, Vienna is the best city in the world to live in, while Baghdad is the worst. London, Paris, and New York are not even in the top 35. German-speaking cities, such as Vienna, Zurich, Munich, Dusseldorf, and Frankfurt, are among the top cities in the ranking. Paris and London have dropped in the rankings, mainly because of the risk of terrorist attacks. The study looks at factors such as social and economic conditions, health, education, housing, and the environment. Companies use this information to decide where to locate and how much to pay their employees. Vienna is particularly good because it has affordable housing, good public transport, and a safe environment. The city has a long tradition of investing in high-quality social housing, which keeps rents low. For example, a 100-square-meter apartment in Vienna costs about 800 euros a month, while a similar apartment in London or New York would cost much more. US cities generally rank lower in the study because of issues with personal safety and crime. San Francisco is the highest-ranking US city at 28th place, followed by Boston at 34th. Canadian cities, such as Vancouver, rank higher than their US counterparts. Vienna is also known for its good quality of life, with good public transport, recreational facilities, and a stable environment. It has a high number of students and is a lively city. The city has been growing in popularity as a tourist destination, especially for British visitors. London, on the other hand, has a poor ranking due to air pollution, traffic congestion, and climate issues. Paris has also dropped in the rankings because of terrorist attacks in 2015. The cities with the worst quality of life are those affected by war and political unrest. Damascus, Baghdad, and cities in countries like Yemen and Haiti are among the worst-ranked. Austrian cities like Vienna have benefited from their location and good economic conditions. They have a high GDP per capita and are good for business. Overall, the study shows that cities with good social and economic conditions, safety, and affordable living are the best places to live.