FOSU_cr at the CLEF 2026 SimpleText Track: A Reference-Guided Few-Shot Prompting Approach for Biomedical Text Simplification

SimpleText at CLEF 2026

Run Chen[1], Yang Yang[2], HuaiYu Zhang[1] and LeiLei Kong[2,*]

[1]Foshan University, GuangDong, China
[2]Center for Disease Control and Prevention of Southern Theatre Command, GuangDong, China

Abstract

The automatic simplification of biomedical scientific texts aims to make specialized research findings accessible to non-expert readers. This paper addresses the CLEF 2026 SimpleText Track Task 1 (biomedical text simplification) and proposes a biomedical text simplification method based on reference-guided few-shot prompting. The method selects authentic Cochrane plain language summaries from the reference distribution of the BioCLEAR benchmark as few-shot examples, and uses in-context learning to guide large language models in generating simplified texts that conform to domain conventions. On the full BioCLEAR test set (666 documents across 5 data sources), the proposed method achieved SARI scores of 46.08 on the `simple_original` reference set and 43.86 on the `simple_auto` reference set for document-level simplification. On the CLEF 2026 official submission evaluation (Cochrane-auto 2026 subset, 6,928 documents), it further achieved SARI scores of 48.85 on the `simple_original` reference set and 47.97 on the `simple_auto` reference set. For sentence-level simplification, the method achieved SARI scores ranging from 41.89 to 44.98 across evaluation subsets. Ablation experiments demonstrate that few-shot examples selected from the target reference distribution are the primary driver of performance improvement, validating the effectiveness of domain-matched examples in biomedical text simplification.

Keywords

Text simplification, Few-shot prompting, Large language model, Cochrane, Plain language summary

1. Introduction

Cochrane systematic review abstracts are key knowledge carriers in evidence-based medicine. Their original texts contain extensive specialized terminology, statistical indicators, and complex sentence structures, which severely hinder non-expert readers from accessing and understanding medical evidence [1, 2]. To address this problem, Task 1 of the CLEF 2026 SimpleText Track [3, 4] uses the BioCLEAR benchmark as its evaluation platform and defines two text simplification subtasks: Task 1.1 (sentence-level simplification), which requires rewriting individual sentences from scientific abstracts into plain language versions; Task 1.2 (document-level simplification), which requires rewriting complete Cochrane systematic review abstracts into plain language summaries for the general public. Both tasks belong to style transfer tasks—given a complex source text X, generate a simplified text Y such that Y preserves the core information of X while reducing lexical complexity and syntactic complexity.

Current biomedical text simplification methods can be categorized into three paradigms. The first is sequence-to-sequence fine-tuning, which fine-tunes pre-trained models such as BART [5] and T5 [6] on complex-simple parallel sentence pairs [7, 8]. These methods achieve strong performance on general simplification benchmarks such as ASSET [9], but their effectiveness heavily depends on large-scale in-domain parallel corpora, and parallel simplification data in the biomedical domain are relatively scarce, limiting the generalization ability of fine-tuning methods. The second is pipeline approaches,

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*Corresponding author.
†These authors contributed equally.
2112571059@stu.fosu.edu.cn (R. Chen); sevenkll@hotmail.com (Y. Yang); zhanghuaiyu2005@gmail.com (H. Zhang); kongleilei@fosu.edu.cn (L. Kong)

which decompose the simplification process into independent steps such as lexical substitution, sentence splitting, and content deletion [10]. Siddharthan [11] pointed out that such modular methods suffer from error propagation across stages, where errors in upstream modules cascade and amplify, affecting the coherence of the final output. The third is prompting approaches based on large language models, which directly generate simplified text through zero-shot or few-shot prompting [12]. Dong et al. [13] systematically surveyed recent advances in in-context learning and found that the selection and arrangement of few-shot examples have a decisive impact on downstream task performance.

However, existing prompting-based methods face a critical bottleneck—the domain style alignment problem: generic simplification prompts cannot capture the specific writing conventions of the target domain. For example, Cochrane plain language summaries organize content with fixed discourse structures such as "What did we want to find out?"; texts generated by zero-shot generic prompts exhibit systematic deviations from such conventions.

To address the above problems, this paper proposes a two-component biomedical text simplification method. Reference-guided refers to the strategy of selecting few-shot examples from the target task's reference distribution. Unlike traditional random selection or similarity-based retrieval methods, reference-guided sampling draws directly from the target distribution, making examples naturally consistent with the expected output at three levels: lexical choice, output format, and style conventions. The core innovation of this strategy is that it bypasses the limitations of general corpora and obtains examples directly from the target task's reference distribution, thereby ensuring that examples are highly aligned with the target output in terms of domain terminology, discourse structure, and writing style. Compared to randomly selected examples, reference-guided examples provide more precise domain signals; compared to similarity-based retrieval examples, reference-guided examples are more consistent with the target distribution in output format and style conventions. In implementation, reference-guided examples are added during prompt template construction. Specifically, this paper selects 3 authentic Cochrane plain language summaries from the `simple_original` reference set of the BioCLEAR benchmark based on length constraints, and embeds them as examples in the system prompt to construct a domain-anchored prompt template. Unlike generic simplification prompts or manually constructed examples, these examples come from the target reference distribution and contain authentic domain terminology mappings (e.g., replacing "intervention" with "treatment") and discourse structures (e.g., question-and-answer paragraph organization), enabling the model to reproduce the target-style lexical choices and syntactic simplification patterns during inference.

2. Related Work

2.1. Biomedical Text Simplification

Automatic simplification of biomedical texts targets professional literature such as Cochrane systematic review abstracts. The core challenges include dense specialized terminology, statistical indicators embedded in sentence structures, and the accuracy constraint that medical facts must not be lost during simplification (Shardlow 2014). The BioCLEAR benchmark (Schafer et al. 2026) provides a standardized evaluation framework for this task, including two types of gold standards: manually written `simple_original` references and automatically aligned `simple_auto` references. Existing biomedical text simplification methods can be categorized into three paradigms: sequence-to-sequence fine-tuning (Martin et al. 2020; Cooper and Shardlow 2020), pipeline approaches (Qiang et al. 2020), and prompting approaches based on large language models (Kew et al. 2023). The comparative experiments of Kew et al. (Kew et al. 2023) show that LLM-based prompting approaches outperform fine-tuned BART and T5 models on BioCLEAR: GPT-4's SARI scores fall in the 38–45 range, while fine-tuned models exhibit significant performance degradation on cross-domain subsets. This finding establishes LLM prompting as the dominant paradigm for biomedical simplification.

2.2. Exemplar Selection in In-Context Learning

In-context learning (ICL) enables LLMs to adapt to downstream tasks by embedding a small number of input-output examples in the prompt, without parameter updates [14]. Dong et al. [13] categorized ICL into three stages—exemplar design, exemplar ordering, and exemplar reasoning—and identified exemplar design as the key factor determining the performance upper bound. Regarding how to select examples, Lu et al. [15] found that the ordering of few-shot examples can cause performance fluctuations of up to 20 percentage points, indicating that models are highly sensitive to exemplar signals. Liu et al. [16] further categorized exemplar selection strategies into three categories—random selection, input similarity-based retrieval, and direct sampling from the target reference distribution—and systematically validated the superiority of the third strategy across multiple text generation tasks: examples from the target distribution are naturally consistent with the expected output at three levels—lexical choice, output format, and style conventions—and provide more complete signals than retrieval-based examples.

However, existing ICL research in text simplification [12] primarily uses manually constructed or randomly selected examples, and has not systematically explored the impact of target reference distribution examples on domain-specific simplification tasks. This paper transfers the findings of Liu et al. [16] from general text generation to the biomedical simplification scenario, directly using authentic Cochrane plain language summaries from the BioCLEAR reference set as few-shot examples, anchoring the in-context signal to the task's target distribution.

3. Method

3.1. System Overview

The proposed system consists of two cascaded components: a reference-guided few-shot prompting module and a rule-based statistical information removal module. Figure 1 illustrates the overall system pipeline. The input is a Cochrane systematic review abstract, which is first fed through a few-shot prompt template to the DeepSeek-V3 model to generate initial simplified text, then processed by a regular expression post-processing step to remove residual statistical symbols, and finally output as a plain language summary.

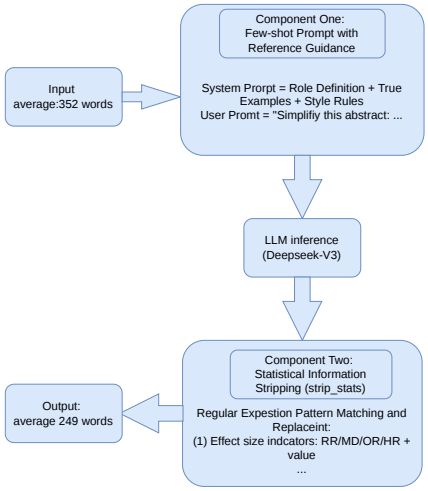

Figure 1: Overall system pipeline

3.2. Reference-Guided Few-Shot Prompting

The selection of few-shot examples directly affects the quality of in-context learning. This paper extracts examples from the BioCLEAR evaluation set that satisfy the following constraints: (1) the reference data contain a `simple_original` field (i.e., manually written plain language summaries officially published by Cochrane); (2) the source text length is between 200 and 400 words; (3) the reference summary length is between 150 and 350 words. The length constraints ensure that the selected examples are representative in text scale for typical documents in the evaluation set. The selection process iterates over all `pair_id` entries in the BioCLEAR reference file, takes the first 3 examples that satisfy the constraints, and embeds their source texts and reference texts as pairs into the system prompt.

The system prompt follows a three-part structure. The first part is the role definition: "You are a Cochrane plain language summary writer." The second part contains 3 examples, each formatted as "Example N - Scientific text: {source}\n\nExample N - Plain language version: {reference}." The third part specifies style rules, requiring the model to simplify according to the length and style of the examples, remove all statistical information, and output 200–300 words. The user prompt uses a fixed template: "Simplify this abstract:\n\n{input_abstract}\n\nPlain language version:".

For sentence-level simplification (Task 1.1), the system prompt remains the same as the document-level one, ensuring consistency of style conventions. The user prompt is changed to a batch processing format: each batch of 15 sentences is input in numbered form (`[1] sentence1\n[2] sentence2\n...`), and the model is asked to return simplified results line by line in the same numbered format. This batch strategy maintains sentence-level alignment relationships while allowing sentences within a batch to share the same contextual style signal.

3.3. Statistical Information Removal

Although the few-shot prompt already contains explicit statistical information removal instructions, residual statistical symbols frequently appear in LLM outputs. Cochrane originals are densely populated with statistical indicators such as risk ratios (RR), confidence intervals (95% CI), p-values, and I-squared. Even when LLMs receive explicit removal instructions, they frequently retain such symbols in their outputs. The empirical study of Liu et al. (Liu et al. 2022) demonstrated that instruction strength in prompts exhibits a decay effect: model compliance with fine-grained constraints (e.g., "delete all statistical numbers") is significantly lower than compliance with high-level semantic instructions (e.g., "rewrite in simple language"). To address this problem, this paper designs a rule-based post-processing module that applies a sequence of regular expressions to remove residual effect size indicators, confidence intervals, p-values, heterogeneity metrics, and certainty of evidence labels from LLM outputs, and cleans up resulting empty parentheses and extra spaces.

1. Bracketed effect size and confidence interval combinations: matches complete statistical reports of the form `(RR 1.73, 95% CI 0.83 to 3.61)`.
2. Standalone effect size indicators: matches effect size expressions such as `RR 1.73`, `MD -0.45`, `OR 2.10`, and `HR 0.85` (i.e., RR/MD/OR/HR followed by a decimal number, including negative values).
3. Confidence intervals: matches confidence interval expressions such as `95% CI 0.83 to 3.61` (i.e., a percentage followed by `CI` and two decimal numbers separated by `to`).
4. p-values: matches significance indicators such as `P = 0.03`, `p < 0.001`, `p > 0.05`, and parenthesized forms `(p = 0.03)`.
5. Heterogeneity metrics: matches heterogeneity expressions such as `I-squared = 45%`.
6. Certainty of evidence labels: matches "very low-certainty evidence," "low-certainty evidence," "moderate-certainty evidence," and "high-certainty evidence."
7. Residual cleanup:
    - Removes empty parentheses `( )` and `()` produced after matching.
    - Merges consecutive spaces (two or more) into a single space.

- Removes extra spaces before punctuation marks.
- Merges consecutive blank lines.

All patterns are executed sequentially using compiled regular expression objects from the Python `re` module in a case-insensitive manner. The patterns are applied from top to bottom: compound patterns with more context (e.g., complete statistical reports within brackets) are matched first, followed by independent numerical patterns, and finally space and punctuation cleanup. This ordering prevents independent patterns from prematurely deleting values, which would cause compound pattern matching to fail.

4. Experiments

4.1. Datasets and Evaluation Metrics

Experiments are conducted at two evaluation levels: (1) BioCLEAR local evaluation, using the officially released evaluation script `evaluate.py`, covering 666 documents (5 data sources); (2) CLEF 2026 official submission evaluation, submitting the complete system-generated output to the competition platform, evaluated on the Cochrane-auto 2026 subset (6,928 unique documents).

The 666 documents in the BioCLEAR local evaluation set come from five data sources:

Table 1
Data source composition of the BioCLEAR evaluation set

Source	#Docs	Description
Cochrane	217	CLEF 2025 SimpleText test data with original Cochrane abstracts
Cochrane-auto val	119	EMNLP/TSAR 2024 Cochrane-auto validation set
Cochrane-auto test	117	EMNLP/TSAR 2024 Cochrane-auto test set
Medline	110	Medline abstract data from TREC PLABA Track
SimpleText2024	103	Manually simplified Arminer abstracts from CLEF SimpleText 2024

Each document provides one or two reference simplifications: (1) simple_original (453 documents): manually written plain language summaries officially published by Cochrane, representing the gold standard for target simplification style; (2) simple_auto (273 documents): sentence-level simplifications extracted from source texts via automatic alignment, retaining only source sentences that are highly aligned with reference sentences. The average length of source abstracts is 352.3 words.

For sentence-level evaluation, BioCLEAR provides 5,384 sentence pairs with `simple_auto` references, covering four sources: Cochrane (363 sentences), Cochrane-auto val (1,041 sentences), Cochrane-auto test (932 sentences), and Medline (709 sentences). Each source sentence in the Medline subset is paired with three manually written simplification references (adaptation1, adaptation2, adaptation3).

The evaluation metric is SARI [17], computed as the arithmetic mean of three sub-score F1 values at the n-gram (1- to 4-gram) level:

$$\text{SARI} = \frac{\text{ADD}_{\text{F1}} + \text{KEEP}_{\text{F1}} + \text{DELETE}_{\text{F1}}}{3}$$

where ADD measures the degree to which n-grams added by the system match n-grams added in the reference, KEEP measures the degree to which n-grams shared between the source text and the reference are retained in the system output, and DELETE measures the proportion of n-grams correctly deleted by the system. SARI ranges from 0 to 100, with higher scores indicating better simplification quality. Evaluation uses the `corpus_sari` function implemented in the EASSE library [9], using the macro-SARI variant (F1 at each n-gram level is computed first and then averaged).

4.2. Experimental Setup

Model. We use DeepSeek-V3 (`deepseek-chat`), accessed via an OpenAI-compatible API endpoint. DeepSeek-V3 is a large language model based on the Mixture-of-Experts (MoE) architecture, with a total of 671B parameters and 37B activated parameters per token [18].

Hyperparameters. The temperature is set to 0.3 to balance between generation diversity and determinism. The maximum output token count is set to 4,096. API call concurrency is 5. The retry strategy allows up to 3 attempts with exponential backoff (wait intervals of 2s, 4s, 6s). The batch size for sentence-level simplification is 15 sentences/request, with a 0.3-second interval between batches to prevent API rate limiting.

Few-shot exemplar configuration. 3 authentic Cochrane plain language summaries from the BioCLEAR `simple_original` reference set, with selection constraints of 200–400 words for the source text and 150–350 words for the reference summary.

Baseline. GPT-4o is the reference system output provided by BioCLEAR (`runs/Test_task12_gpt4o.json` and `runs/Test_task11_gpt4o.json`), which uses GPT-4o to generate simplified texts for the BioCLEAR evaluation set, representing the official baseline of the CLEF 2026 SimpleText Track. All evaluations in this section are executed on the official BioCLEAR evaluation set (`sources/abstracts.json`, 666 documents) using the `evaluate.py` script, with input sources and reference answers identical to the baseline system.

To quantify the independent contribution of each component, we define the following ablation configurations (see Section 4.4 for details):

- Zero-shot Prompt: DeepSeek-V3 uses a generic simplification prompt with no few-shot examples and no post-processing.
- + Domain Prompt + PostProc: adds statistical removal instructions and lexical replacement guidelines to the Zero-shot Prompt, and enables the post-processing module.
- + Reference Exemplars (Ours): + Domain Prompt + PostProc + reference-guided few-shot examples.

4.3. Main Results

Document-level results. Table 2 shows the SARI scores of each method on the BioCLEAR document-level simplification task.

Table 2
Document-level SARI scores (BioCLEAR evaluation set, 666 documents)

Method	Cochrane	Coch.-auto val	Coch.-auto test	Medline	SimpleText2024
GPT-4o (official baseline)	42.21	38.90	39.04	37.36	42.68
Ours	**47.16**	**44.79**	**43.20**	**41.99**	**44.24**

Note: Cochrane-series data sources use `simple_original` as the reference; Medline uses the adaptation reference set; SimpleText2024 uses `simple` as the reference. The proposed method consistently outperforms the GPT-4o baseline across all data sources, with the largest improvement on the Cochrane subset (+4.95 points).

Table 3
Document-level SARI sub-scores (`simple_original` reference)

Method	ADD	KEEP	DELETE	SARI
GPT-4o (official baseline)	6.59	27.19	89.07	40.95
Ours	**10.49**	**40.05**	**87.69**	**46.08**

To further analyze the contribution of each sub-score, Table 3 lists the ADD, KEEP, and DELETE sub-scores on the `simple_original` reference set.

GPT-4o's DELETE score is as high as 89.07, indicating a strong tendency to delete statistical information, but its KEEP score is only 27.19, reflecting excessive deletion of medical content that should have been retained. The proposed method's ADD score improves by 3.90 points, and its KEEP score substantially improves by 12.86 points, while the DELETE score remains at a high level of 87.69. The improvement in KEEP score is the primary contributing factor to the overall SARI improvement, indicating that reference-guided few-shot examples enable the model to more accurately identify core semantic content that should be retained.

Sentence-level results. Table 4 shows the sentence-level SARI scores for each evaluation subset.

Table 4
Sentence-level SARI scores (BioCLEAR evaluation set, 5,384 sentences)

Method	Cochrane	Coch.-auto val	Coch.-auto test	Medline	Average
GPT-4o (official baseline)	38.58	36.43	36.27	39.43	37.68
Ours	**44.98**	**43.64**	**41.89**	**43.47**	**43.50**

The proposed method achieves consistent improvements over the GPT-4o baseline across all subsets (+5.59 to +7.28). The Cochrane subset shows the largest improvement (+6.40), where the reference is the official manually written plain language summary. The Cochrane-auto test and val subsets have lower GPT-4o scores (36.27–36.43) because the references for these subsets are automatically aligned texts, and GPT-4o's generic simplification output has lower n-gram matching degree with the aligned references. The performance gap between sentence-level and document-level (document-level average 44.73 vs. sentence-level average 43.50) indicates that paragraph-level contextual information helps the model more accurately determine which content should be retained and which should be deleted.

Table 5 lists the SARI sub-scores on the `simple_auto` reference.

Table 5
Sentence-level SARI sub-scores (`simple_auto` reference, 2,336 sentences)

Method	ADD	KEEP	DELETE	SARI
GPT-4o (official baseline)	4.06	28.00	78.12	36.73
Ours	**5.66**	**46.38**	**76.24**	**42.76**

Consistent with the document-level trend, the KEEP score shows the largest improvement (+18.38), validating that reference-guided few-shot examples equally help the model identify core semantic content to retain at the sentence level. The sentence-level DELETE score (76.24) is lower than the document-level score (87.69) because individual sentences lack paragraph-level context, and the model's confidence in determining whether a given n-gram should be deleted is lower than at the document level. GPT-4o at the sentence level also exhibits the high DELETE, low KEEP pattern (DELETE 78.12, KEEP 28.00), consistent with the document-level analysis.

Official submission evaluation results. The proposed method was applied to the full Task 1.2 input data (6,928 unique documents) to generate `documents_dst.json`, which was submitted to the CLEF 2026 official evaluation platform. The platform computed SARI and auxiliary metrics on the Cochrane-auto 2026 subset using `simple_original` and `simple_auto` as references, respectively. Table 6 shows the complete evaluation results returned by the platform.

The official evaluation SARI scores (48.85/47.97) are higher than the BioCLEAR local evaluation scores (46.08/43.86). The difference arises from variations in evaluation scope: the official evaluation covers only the single Cochrane-auto 2026 source, where the source texts have more uniform structures and statistical information densities, making simplification difficulty relatively concentrated; whereas the BioCLEAR local evaluation additionally covers cross-domain sources such as Medline and SimpleText2024, and the diversity of text characteristics increases simplification difficulty.

Table 6
Official submission evaluation results (Cochrane-auto 2026 subset, 6,928 documents)

Metric	`simple_original` ref.	`simple_auto` ref.
SARI	48.85	47.97
BLEU	12.22	16.96
FKGL	8.80	8.54
Compression ratio	0.522	0.551
Sentence split ratio	0.878	0.975
Levenshtein similarity	0.453	0.476
Exact copy rate	0.0	0.0
Addition ratio	0.178	0.182
Deletion ratio	0.659	0.635
Lexical complexity	8.47	8.47

The auxiliary metrics further characterize the simplification output: (1) the exact copy rate is 0.0, indicating that all output texts underwent substantial rewriting; (2) the deletion ratio is as high as 0.635–0.659, reflecting the effectiveness of statistical information removal; (3) the BLEU scores are low (12.22–16.96), consistent with prior findings that BLEU has low correlation with human evaluation in simplification tasks [17, 9].

4.4. Ablation Study

To quantify the independent contribution of each component, an ablation study was conducted on the `simple_original` reference set. Starting from the most basic generic prompt configuration, the effects of incrementally adding each component are shown in Table 7.

Table 7
Ablation study (`simple_original` reference)

Configuration	ADD	KEEP	DELETE	SARI	Δ SARI
Zero-shot Prompt	7.06	31.22	88.34	42.20	–
+ Domain Prompt + PostProc	7.84	32.74	88.80	43.12	+0.92
+ Generic Exemplars	7.65	29.23	89.17	42.02	−1.10
+ **Reference Exemplars (Ours)**	**10.49**	**40.05**	**87.69**	**46.08**	**+3.88**

The contribution of each component is analyzed as follows:

(1) Domain Prompt and PostProc (Zero-shot Prompt → + Domain Prompt + PostProc, +0.92 SARI): Enhanced statistical deletion instructions and terminology replacement guidelines in the prompt lead to minor improvements in all three sub-scores (ADD, KEEP, DELETE), indicating that domain-specific prompt engineering can produce positive gains. However, the improvement is limited (SARI increases by only 0.92), suggesting that text instructions alone cannot fundamentally change the model's generation behavior.

(2) Generic Exemplars (+ Domain Prompt + PostProc → + Generic Exemplars, −1.10 SARI): After introducing 3 non-biomedical-domain simplification examples, SARI drops from 43.12 to 42.02. The KEEP score sharply decreases from 32.74 to 29.23 (−3.51), indicating that general-domain examples introduce incorrect discourse structure signals, causing the model to discard medical content that should have been retained during simplification. This result is consistent with the conclusion of Liu et al. [16] that exemplar selection determines the ICL performance upper bound: examples from mismatched distributions are not only unhelpful but actually cause negative transfer.

(3) Reference Exemplars (+ Generic Exemplars → + Reference Exemplars (Ours), +4.06 SARI): Switching the example source to the `simple_original` reference distribution of BioCLEAR, SARI jumps from 42.02 to 46.08 (+4.06), with ADD and KEEP improving to 10.49 and 40.05, increases of +2.84 and +10.82,

respectively. Terminology replacements in the reference distribution examples (e.g., "intervention" → "treatment") and discourse structures (e.g., question-and-answer paragraph organization) are fully aligned with the target task, enabling the model to accurately reproduce target-style lexical choices and syntactic simplification patterns during inference. The improvements in ADD and KEEP are mutually independent and positively synergistic with the DELETE function of the post-processing module; the two components are complementary rather than competitive.

4.5. Output Feature Analysis

Table 8 compares the statistical characteristics of source texts, reference texts, and the proposed method's output texts.

Table 8
Comparison of output text features

Metric	Source	Reference (`simple_original`)	Our output
Average length (words)	352.3	312.0	249.0
Compression ratio	1.00	0.89	0.71

The average length of the proposed method's output texts is 249 words, with a compression ratio of 0.71, higher than the reference's 0.89. The higher compression degree is jointly caused by two factors: the explicit length constraint of "Write 200–300 words" in the system prompt, and the additional length reduction produced by the post-processing module after removing statistical information. The higher compression ratio helps achieve higher DELETE scores, but may to some extent limit the further improvement potential of the KEEP score—some content compressed by the length constraint should have been retained.

5. Conclusion

Biomedical text simplification is a key task for making evidence-based medicine accessible to the public. This paper proposes a biomedical text simplification method called reference-guided few-shot prompting, aimed at addressing the challenges of existing LLM-based prompting methods: mismatch between few-shot examples and the target distribution leads to systematic deviations in domain style and terminology choice of generated texts. To this end, this paper introduces the reference-guided exemplar selection strategy into the biomedical simplification scenario, constructing few-shot examples by directly sampling from the target reference distribution. The method first selects authentic Cochrane plain language summaries from the `simple_original` reference set of the BioCLEAR benchmark based on length constraints, and embeds them as examples in the system prompt to construct a domain-anchored prompt template, enabling the model to reproduce target-style lexical choices and syntactic simplification patterns during inference. Meanwhile, this paper employs a rule-based post-processing module to address the problem of residual statistical symbols in LLM outputs. The method does not require fine-tuning model parameters and accomplishes domain style alignment and simplified text generation entirely at the prompt level.

Experiments on the BioCLEAR benchmark evaluation set and the CLEF 2026 official submission evaluation validated the effectiveness of the proposed method.

The limitations of this work are as follows: the statistical information removal module is based on hand-crafted regular expressions, and its coverage is limited to predefined patterns; additionally, the few-shot exemplar selection strategy is relatively simple, relying solely on length constraints. Future work can explore more intelligent exemplar selection methods, such as similarity-based retrieval strategies, and more robust statistical information identification and conversion methods.

Acknowledgments

This research was supported by the National Social Science Foundation of China (22BTQ101).

Declaration on Generative AI

During the preparation of this work, the authors used *Gemini* in order to: **improving writing style** and **citation management**. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] M. Shardlow, A survey of automated text simplification, International Journal of Advanced Computer Science and Applications 4 (2014) 58–70. Special Issue on NLP.

[2] L. Ermakova, E. SanJuan, S. Huet, H. Azarbonyad, O. Augereau, J. Kamps, Overview of the CLEF 2023 SimpleText lab: Automatic simplification of scientific texts, in: Proceedings of the 14th International Conference of the CLEF Association (CLEF 2023), volume 14163 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 482–506. doi:10.1007/978-3-031-42448-9_30.

[3] L. Ermakova, E. SanJuan, S. Huet, O. Augereau, H. Azarbonyad, J. Kamps, Overview of CLEF 2026 SimpleText track: Simplify scientific texts (and nothing more), in: M. Hagen, et al. (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Lecture Notes in Computer Science, Springer, 2026.

[4] J. Bakker, B. Vendeville, L. Ermakova, J. Kamps, Overview of the CLEF 2026 SimpleText Task 1: Simplify scientific text, in: E. S. Salido, A. B.-C. no, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2026), CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[5] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020, pp. 7871–7880.

[6] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of Machine Learning Research 21 (2020) 1–67.

[7] L. Martin, Éric de la Clergerie, B. Sagot, A. Bordes, Controllable sentence simplification, in: Proceedings of the 12th Language Resources and Evaluation Conference (LREC), 2020, pp. 4689–4698.

[8] M. Cooper, M. Shardlow, CombiNMT: An exploration into neural text simplification models, in: Proceedings of the 12th Language Resources and Evaluation Conference (LREC), 2020, pp. 5588–5594.

[9] F. Alva-Manchego, L. Martin, A. Bordes, C. Scarton, B. Sagot, L. Specia, ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020, pp. 4668–4679. doi:10.18653/v1/2020.acl-main.424.

[10] J. Qiang, Y. Li, Y. Zhu, Y. Yuan, X. Wu, Lexical simplification with pretrained encoders, in: Proceedings of the 34th AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 8649–8656.

[11] A. Siddharthan, Syntactic simplification and text cohesion, in: Research on Language and Computation, volume 4, Springer, 2006, pp. 77–109.

[12] T. Kew, A. Chi, L. Vasquez-Rodriguez, S. Agrawal, D. Aumiller, F. Alva-Manchego, M. Shardlow, BLESS: Benchmarking large language models on sentence simplification, in: Proceedings of

the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2023, pp. 13291–13309.

[13] Q. Dong, L. Li, D. Dai, C. Zheng, J. Ma, R. Li, H. Xia, J. Xu, Z. Wu, B. Chang, X. Sun, L. Li, Z. Sui, A survey on in-context learning, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2024.

[14] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: Advances in Neural Information Processing Systems 33 (NeurIPS), 2020.

[15] Y. Lu, M. Bartolo, A. Moore, S. Riedel, P. Stenetorp, Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL), 2022, pp. 8086–8098.

[16] J. Liu, D. Shen, Y. Zhang, B. Dolan, L. Carin, W. Chen, What makes good in-context examples for GPT-3?, in: Proceedings of Deep Learning Inside Out (DeeLIO): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures at ACL, 2022, pp. 100–114.

[17] W. Xu, C. Napoles, E. Pavlick, Q. Chen, C. Callison-Burch, Optimizing statistical machine translation for text simplification, Transactions of the Association for Computational Linguistics (TACL) 4 (2016) 401–415.

[18] DeepSeek-AI, DeepSeek-V3 technical report, arXiv preprint arXiv:2412.19437 (2024).