Medical Text Rewriting for Non-Experts: A Guideline-Driven LLM Approach

Mana Kuramoto[1], **Hiroyuki Nagai**[1], **Keiko Yamada**[2], **Hiroo Ide**[3,4]
Masayo Hayakawa[3,5], **Tomohiro Nishiyama**[1], **Shoko Wakamiya**[1], **Eiji Aramaki**[1]
[1]Nara Institute of Science and Technology, Japan
[2]Saitama Prefectural University, Japan [3]The University of Tokyo, Japan
[4]Juntendo University, Japan [5]Keio University, Japan
{kuramoto.mana.kj4, hiro.nagai}@naist.ac.jp, yamada-keiko@spu.ac.jp,
hiroo-ide@g.ecc.u-tokyo.ac.jp, hayakawam395@gmail.com,
{nishiyama.tomohiro.ns5, wakamiya, aramaki}@is.naist.jp

Abstract

Medical research is highly specialized, making it difficult for patients and general readers to understand recent findings. Traditionally, text simplification, replacing technical terms with more accessible expressions, has been employed. However, this approach alone is limited in addressing a lack of background knowledge and often results in the loss of important information. Therefore, this study defines "rewriting for non-experts" as a rewriting process that, in addition to simplification, supplements essential background knowledge such as the significance of the research and reasons it is needed and proposes a method for implementing this process using large language models (LLMs). To verify the effectiveness of the proposed approach, a quantitative evaluation using automatic metrics was conducted. The results showed that the method combining the guidelines for human text creation with few-shot examples of reference texts achieved the highest scores. The expansion of the guidelines is planned as part of future work to enable the rewriting of scientific and technological information in a form that is accessible to a broader audience.

Keywords: Medical Text Rewriting, Large Language Models, Comprehensibility Enhancement, Text Simplification, Lay Summarization, Non-expert Communication

1. Introduction

Scientific specialization has intensified the "silo effect," isolating research fields and hindering cross-disciplinary communication (Tett, 2015). As expertise narrows and modern science becomes increasingly complex, even experts struggle to grasp developments outside their domains. Consequently, cutting-edge research remains largely inaccessible to laypeople. This accessibility gap is particularly critical in medicine: although open science grants patients access to primary literature, interpreting these texts without specialized training remains difficult (Gotlieb et al., 2022). Misinterpreting medical information poses significant risks, potentially disrupting patient–physician communication and compromising clinical decisions (Lu and Schulz, 2024; Root et al., 2016).

Current text simplification methods, which focus primarily on vocabulary substitution, fail to address deeper structural barriers (Alva-Manchego et al., 2021). The difficulty of scientific text stems not just from jargon, but from implicit assumptions, dense structure, and the reader's lack of background knowledge. To bridge this gap, we introduce "rewriting for non-experts"—an advanced editing paradigm that, beyond simple simplification, involves selecting salient information, restructuring the narrative, and clarifying the research's significance. Fig. 1 illustrates this process: rather than merely simplifying vocabulary, the transformation explicates implicit context and foregrounds essential information while retaining necessary technical terms. This demonstrates that enhancing comprehensibility requires structural reorganization and contextual clarification.

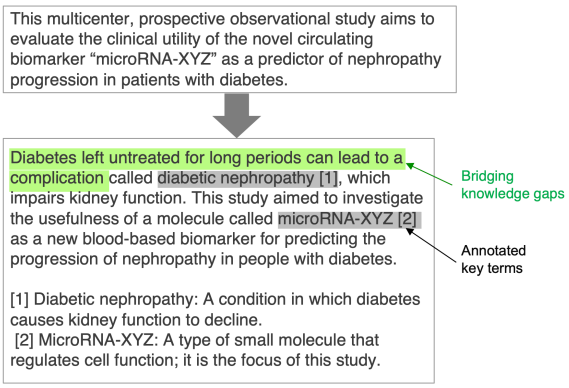

Figure 1: An example of rewriting for non-experts. The original expert-oriented text (top) is transformed into an accessible version (bottom) by explicating implicit context and restructuring the narrative.

Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health), pages 107–116
12 May 2026.

As systematized in Fig. 2, our approach addresses three key barriers: (1) Implicit Knowledge: Unstated premises assumed by specialists. (2) Excessive Statistical Detail: Quantitative data that increases cognitive load without aiding non-expert understanding (Gigerenzer and Edwards, 2003). (3) Disrupted Flow: Frequent definitions that interrupt the narrative coherence (Leroy et al., 2013).

Consider a concrete scenario illustrating these barriers: a patient reads a press release about a promising new cancer drug. The text states that the drug "showed significant efficacy in clinical trials." However, critical context—that the drug remains years from regulatory approval and is unavailable outside research settings—is omitted as self-evident to researchers. Such implicit assumptions can lead patients to request unavailable treatments, straining clinical relationships (Lu and Schulz, 2024; Root et al., 2016). This accessibility gap requires "rewriting for non-experts"—an advanced editing process that goes beyond simplification to include selecting key information, restructuring narratives, and explicating significance (Schmitz, 2023).

This study proposes a comprehensive rewriting framework that prioritizes comprehensibility while preserving essential medical accuracy. To this end, we investigate the feasibility of this complex task using Large Language Models (LLMs). Leveraging their capacity to integrate contextual information across discourse, the proposed approach employs LLMs to perform structural reorganization and context supplementation guided by human-authored references, distinguishing our approach from conventional simplification models.

2. Related Work

2.1. Lay Summarization

Lay summarization is the task of summarizing and restructuring documents containing specialized content into forms that are understandable to non-expert readers. For example, in the medical domain, a generative model that integrates background explanations and lexical substitutions has been proposed for producing summaries suitable for lay audiences (Guo et al., 2021). Furthermore, in the shared task BioLaySumm, a corpus aligning technical abstracts with lay summaries has been developed, along with a continuous evaluation framework (Goldsack et al., 2022, 2024).

Unlike conventional summarization, these studies are notable for aiming to generate documents tailored to the assumed knowledge levels and information needs of target readers. However, they face limitations, such as a narrow target audience and the absence of explicit modeling for appropriately conveying the social significance and novelty of the research. For instance, in frameworks such as BioLaySumm and Explain-type LLMs, while readability and factual consistency are prioritized as evaluation criteria, the question of what kind of value is conveyed to whom is not explicitly addressed (Goldsack et al., 2024; Luo et al., 2024). Moreover, it has been reported that automatic evaluation metrics do not necessarily align with non-experts' judgments of understandability and acceptability (Salvi et al., 2025), highlighting the need to redesign evaluation frameworks that consider reader attributes and perceptions.

2.2. Text Simplification

Text simplification is a task aimed at reducing readers' cognitive load and enhancing comprehensibility by adjusting vocabulary and sentence structures in specialized or complex documents. In recent years, research has advanced across various domains, including news articles, academic writing, and medical texts. Automatic evaluation metrics play a central role in this field. SARI (Xu et al., 2016) is notable for its ability to independently assess additions, deletions, and retentions of lexical items. In contrast, metrics such as BLEU and FKGL (Kincaid et al., 1975) have been reported to exhibit important limitations for simplification evaluation, particularly in capturing simplicity and structural transformations (Alva-Manchego et al., 2021).

Applying these evaluation frameworks to the medical domain presents additional challenges. Medical documents rely heavily on technical terminology and domain-specific concepts, and even minor distortions may affect factual accuracy and safety. Consequently, simplification quality cannot be assessed solely through conventional readability-oriented metrics. Moreover, systematically aligned parallel corpora consisting of expert-level medical texts and their simplified counterparts remain limited. This scarcity complicates validation of automatic evaluation metrics in medical contexts, where meaning preservation and factual consistency are especially critical for rewriting for non-experts (Ondov et al., 2022).

Beyond domain-specific concerns, research has explored models that treat output difficulty as a controllable parameter (Martin et al., 2020), as well as controlled text simplification methods that explicitly specify target reader groups such as grade levels (Scarton and Specia, 2018). These approaches extend uniform simplification strategies by incorporating reader characteristics into the generation process. However, adjusting surface-level textual complexity does not necessarily resolve gaps in background knowledge between experts and non-experts. Empirical evaluations

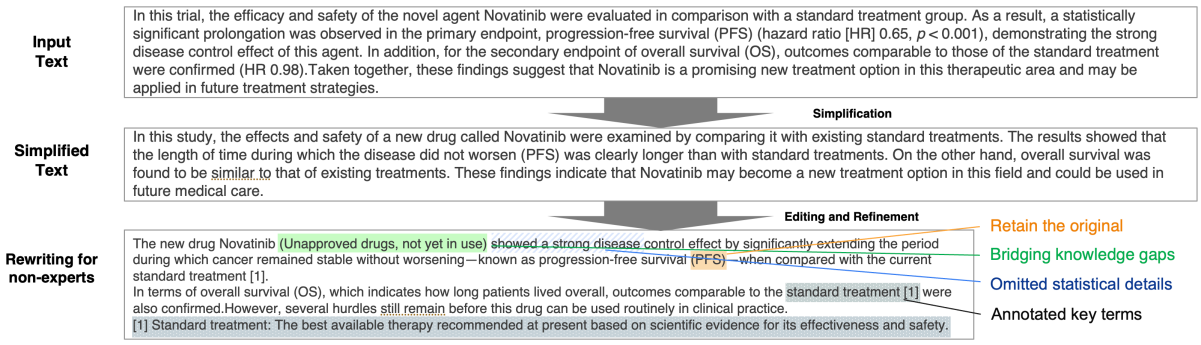

Figure 2: Overview of rewriting operations: supplementing implicit context (green), removing excessive details (blue), and explaining technical terms (orange).

based on reading comprehension tasks further indicate that important information may still be omitted during simplification (Agrawal and Carpuat, 2024). Taken together, these observations suggest that readability alone may not be sufficient to ensure adequate understanding of research background or significance.

2.3. Readability

Readability refers to quantitative measures of how easily a text can be processed by readers, and it has been widely used to evaluate the effectiveness of text simplification and lay summarization. Typical metrics rely on surface-level features such as lexical frequency, sentence length, and syntactic complexity (Kincaid et al., 1975). With the development of large-scale corpora for readability assessment, automatic metrics based on statistical modeling and machine learning have been proposed and applied to compare outputs of generative models (Crossley et al., 2022).

Although readability metrics provide useful indicators of linguistic difficulty, they primarily capture surface characteristics of texts. They do not directly account for how readers interpret information or construct meaning. In other words, even when lexical and syntactic complexity are reduced, abstract or conceptually dense content may remain difficult for readers without relevant background knowledge to understand. In rewriting for non-experts, therefore, formal readability constitutes a necessary condition, but it is not sufficient to ensure effective comprehension.

Recent studies have attempted to extend readability-oriented approaches by incorporating contextual generation processes. For example, multi-stage frameworks have been proposed in which large language models first generate background explanations and subsequently perform summarization (Luo et al., 2024). These approaches aim to enhance accessibility by supplementing contextual information while maintaining linguistic simplicity.

Nevertheless, existing methods primarily focus on adding background information or adjusting textual difficulty at the surface level. They do not systematically reorganize information according to reader attributes or explicitly foreground research novelty and social significance. Thus, while readability remains a fundamental component, a more comprehensive framework that integrates document-level restructuring, contextual supplementation, and audience-sensitive organization remains underexplored. Within this landscape, rewriting for non-experts is positioned as a task that extends readability by emphasizing contextual reconstruction and explicit articulation of research significance for non-expert readers.

3. Dataset

This paper constructs a dataset consisting of pairs of simulated medical research summaries (hereafter, simulated summaries) and reference summaries manually rewritten by a medical writer. Because the copyright of the original reports belongs to individual researchers, we cannot directly release the original texts as a public dataset. To address this limitation and enable public release, we generated fictional summary documents (i.e., simulated summaries) using GPT-4.1-mini to mimic the content and writing style of the source reports.

To construct these simulated summaries, we used research outcome reports from AMEDfind[1], provided by the Japan Agency for Medical Research and Development (AMED), as source material. These simulated summaries focus on neoplasm-related clinical trials and drug development studies, which represent the most frequent category in the source data, and are used as the experimental source texts. The resulting pairs of simulated and reference summaries serve as the

[1] https://amedfind.amed.go.jp/amed/

Item	Description
1. Overall policy	Target audience: Non-medical adult readers without prior knowledge of biology, equivalent to graduates of four-year undergraduate programs in the humanities and social sciences. Wordings should be selected so that readers without specialized knowledge can understand the research content and its key outcomes.
2. Structural and stylistic rules	The original writing style (plain or polite form) should be preserved. Reordering of words or phrases is allowed when necessary to improve clarity. Expressions or sentences that are unnecessary for understanding, such as sample size details or procedural descriptions, should be removed. When information is uncertain, definitive expressions should be avoided and appropriate hedging should be maintained.
3. Terminology rules	A plain expression should be presented first, followed by the original technical term in parentheses. Frequently used terms should be explained using footnotes. Abbreviations should be spelled out in full at their first occurrence. Chained possessives (e.g., "A of B of C") should be avoided by rephrasing or by adding explanatory footnotes.
4. Additional notes	Detailed examination procedures and analysis methods should be summarized concisely and described in plain language whenever possible.

Table 1: Guidelines for rewriting for non-experts

source and target texts for the two-stage rewriting method.

Manual rewriting for comprehensibility was then applied to the simulated summaries. Specifically, 20 documents were selected from the generated corpus, and one professional medical writer was commissioned to perform rewriting for non-experts. Since the simulated summaries were generated by LLMs, they contain no personally identifiable information or actual patient data. The resulting twenty rewritten documents were used as the reference summaries for evaluation.

4. Proposed Method

This paper proposes a two-stage LLM-based method for rewriting for non-experts. The method consists of sequential simplification and refinement stages.

In the Base stage, the LLM performs basic rewriting using prompts that instruct either simple simplification or direct rewriting for non-experts. In the Refinement stage, the output is further refined through prompts guided by predefined guidelines and few-shot examples, enabling more detailed restructuring and contextual adjustment. The two-stage configuration allows systematic comparison between surface-level simplification and guideline-informed refinement.

Base stage In the Base stage, two prompt types are employed to generate base texts for subsequent refinement.

Surface-level Simplification (Simp) : An instruction that focuses exclusively on reducing linguistic complexity at the lexical and syntactic levels, without addressing contextual or background knowledge gaps. Example: *Please simplify the following text so that it can be understood by people without specialized knowledge.*

Comprehensive Rewriting (Direct) : An instruction that directly specifies rewriting for non-experts as the objective, which may involve not only simplification but also selection of key information, structural reorganization, and contextual clarification. The LLM is asked to produce a version appropriate for non-experts, allowing the model to determine the full range of strategies needed to achieve comprehensibility. Example: *Please perform rewriting for non-experts on the following text.*

Refinement stage The output generated in the Base stage is further refined by simulating the human rewriting process. In the proposed method, the prompt in this stage combines the following two components, both of which are applied together to guide the rewriting process.

Guidelines (GL) : A set of editing requirements independently designed and validated by experts. The detailed criteria are summarized in Table 2.3. These guidelines include rules for information supplementation and text structuring, such as adding explanations to technical terms and annotating frequently used terms outside the main text.

Few-shot examples (FS) : Example pairs consisting of original simulated summaries and

Input (Original)	非小細胞肺がん（NSCLC）において，EGFR 変異陽性例を対象に，既存治療薬に耐性を示す症例に対する新規の第三世代 EGFR チロシンキナーゼ阻害剤（TKI）の有効性および安全性評価を目的とした多施設共同第 II 相臨床試験を実施する． A multicenter phase II clinical trial was conducted to evaluate the efficacy and safety of a novel third-generation EGFR tyrosine kinase inhibitor (TKI) in patients with non-small cell lung cancer (NSCLC) who were positive for EGFR mutations and had developed resistance to existing therapies.
Simp	非小細胞肺がんのうち，特定の遺伝子変異がある患者を対象に，これまでの薬が効かなくなった場合の新しい薬の効果と安全性を調べる臨床試験を，複数の施設で行いました． Among patients with non-small cell lung cancer, a clinical trial was conducted at multiple institutions to examine the effectiveness and safety of a new drug in patients with specific gene mutations for whom previous treatments had become ineffective.
Simp+GLFS	非小細胞肺がん [1] の患者さんの中には，特定の遺伝子 [2] に変化がある方がいます．これまでの薬が効かなくなってしまったそのような患者さんを対象に，新しい薬の効き目と安全性を調べる臨床試験を，複数の病院で行いました． [1] 非小細胞肺がん：肺がんの一種で，比較的ゆっくりと進行するタイプです． [2] 遺伝子：体の設計図のようなはたらきをする，細胞の中の物質です． Among patients with non-small cell lung cancer [1], some have changes in specific genes [2]. In such patients for whom previous treatments had become ineffective, a multicenter clinical trial was conducted to evaluate the effectiveness and safety of a new drug. [1] Non-small cell lung cancer : A type of lung cancer that generally progresses relatively slowly. [2] Gene : A substance inside cells that functions like a blueprint for the body.
GLFS	肺がんの一種である非小細胞肺がんの患者のうち，特定の遺伝子（EGFR[1]）に変異がある人を対象に，臨床試験を行った．（中略）新しい第三世代の EGFR チロシンキナーゼ阻害剤（TKI）[2] という薬の効き目と安全性を評価することが目的だった． [1] EGFR：細胞の増殖に関わるタンパク質（受容体）の一種．この遺伝子に変異があると，がん細胞が増えやすくなる． [2] EGFR チロシンキナーゼ阻害剤（TKI）：EGFR というタンパク質のはたらきを抑えることで，がん細胞の増殖を止める薬． A clinical trial was conducted in patients with non-small cell lung cancer, a type of lung cancer, who had mutations in a specific gene, EGFR [1]. ... The objective was to evaluate the efficacy and safety of a new third-generation EGFR tyrosine kinase inhibitor (TKI) [2]. [1] EGFR: A type of receptor protein involved in cell growth; mutations in this gene can promote the proliferation of cancer cells. [2] EGFR tyrosine kinase inhibitor (TKI): A drug that suppresses cancer cell growth by inhibiting the activity of the EGFR protein.

Table 2: Comparison of generation examples produced by DeepSeek-V3.2. Note that English translations are provided for the generated Japanese text. Under the two-stage setting, important technical terms such as EGFR are omitted, whereas the single-stage setting preserves these terms while providing additional explanations.

their corresponding reference summaries are provided to illustrate how the guidelines are applied in practice. Through in-context learning (ICL) , the LLM acquires concrete patterns for generating text consistent with rewriting for non-experts.

5. Experiments

To evaluate the effectiveness of the proposed two-stage rewriting method, comparative experiments using LLMs were conducted. Simulated summary documents were used as input, and the rewritten texts were quantitatively evaluated based on their similarity to the reference summaries produced by a medical writer.

The experiments analyze the impact of two variables on output quality: (1) the type of instruction used in the Base stage, namely simplification or direct rewriting for non-experts, and (2) the presence or absence of refinement using additional information, namely guidelines and few-shot examples, in the Refinement stage.

All texts generated under each experimental setting are publicly available.[2]

[2] https://github.com/sociocom/E2U

Model	Setting		Similarity Metrics		Summary Metrics		
	Base	Refinement	BERTScore	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
DeepSeek	Simp	–	**0.774**	16.827	0.543	0.244	**0.401**
	Direct	–	0.766	15.394	0.563	0.240	0.382
	Original	GL	0.747	11.145	0.742	0.327	0.311
	Original	GLFS	0.773	**19.607**	**0.763**	**0.403**	0.398
	Direct	GL	0.745	12.764	0.641	0.256	0.339
	Direct	GLFS	0.760	15.148	0.627	0.270	0.360
	Simp	GL	0.748	12.743	0.647	0.266	0.337
	Simp	GLFS	0.762	16.917	0.667	0.308	0.375
Gemini	Simp	–	**0.771**	**17.540**	0.633	0.282	**0.393**
	Direct	–	0.762	15.407	0.648	0.276	0.376
	Original	GL	0.730	8.894	**0.777**	**0.361**	0.251
	Original	GLFS	0.753	13.499	0.476	0.248	0.318
	Direct	GL	0.729	9.161	0.727	0.303	0.280
	Direct	GLFS	0.748	12.600	0.722	0.315	0.333
	Simp	GL	0.755	11.079	0.728	0.325	0.296
	Simp	GLFS	0.749	13.023	0.738	0.346	0.321
GPT-5.2	Simp	–	**0.784**	**19.288**	0.591	0.276	**0.420**
	Direct	–	0.777	18.596	0.617	0.286	0.402
	Original	GL	0.750	12.279	0.750	0.347	0.309
	Original	GLFS	0.753	13.060	**0.796**	**0.415**	0.307
	Direct	GL	0.747	13.655	0.705	0.313	0.329
	Direct	GLFS	0.754	15.040	0.704	0.324	0.356
	Simp	GL	0.753	13.123	0.711	0.320	0.325
	Simp	GLFS	0.755	14.322	0.728	0.338	0.343

Table 3: Automatic evaluation results under various rewriting settings. For each model, the best value for each evaluation metric is shown in bold. The shaded rows indicate the proposed two-stage method (Simp+GLFS).

5.1. Models

The following three LLMs were selected for the experiments based on their demonstrated effectiveness in prior text generation and simplification studies:

- **DeepSeek-V3.2** (DeepSeek-AI, 2025): A reasoning-focused language model developed by DeepSeek. Released with open weights and known for cost efficiency.
- **Gemini 2.5 Flash** (Gemini Team, 2025): A multimodal language model developed by Google. Optimized for efficiency and response speed.
- **GPT-5.2 (gpt-5.2-2025-12-11)** (OpenAI, 2025): A general-purpose language model developed by OpenAI. Widely used in natural language processing research.

For parameter settings, the temperature was set to 0.0 and the maximum output length to 2048 tokens for DeepSeek-V3.2 to ensure stable generation. For Gemini 2.5 Flash and GPT-5.2, the default parameter settings provided by each API were used without further modification.

5.2. Evaluation Data and Metrics

For evaluation, 20 pairs of simulated summary documents and their corresponding reference summaries were used, as described in Section 3. In this experiment, two randomly selected pairs were used as few-shot examples in the Refinement stage, and the remaining 18 pairs were used as test data.

To assess the quality of the generated texts from multiple perspectives, **BERTScore** was used to measure semantic similarity, **BLEU** to assess surface-level overlap, and **ROUGE-1/2/L** to evaluate n-gram and sequence-level agreement.

5.3. Ablation Settings

To examine the contribution of each component of the proposed method to rewriting performance, ablation experiments were conducted by varying the combination of instructions in the Base stage (Simp/Direct) and the additional information provided in the Refinement stage (GL/GLFS).

Input	A newly identified transcription factor, XZF-1, has been shown by recent basic research to play an important role in the development and progression of lung cancer. (omitted) Furthermore, based on the design of early-phase clinical trials to evaluate the efficacy of XZF-1 inhibitors, protocols are being developed to establish a foundation for their future introduction as standard treatment. 新規に同定された転写因子 XZF-1 は，肺がんの発生および進展に重要な役割を果たすことが近年の基礎研究で示された．（中略）さらに，XZF-1 阻害剤の有効性を評価する初期臨床試験設計に基づき，将来的な標準治療への導入に向けたプロトコール策定を進めている．
GPT-5.2	A regulatory factor called “XZF-1,” which controls gene activity, has recently been shown through experimental research to be involved in the processes by which lung cancer develops and progresses. (omitted) Based on these findings, trial protocols are being prepared to evaluate the effects of XZF-1 inhibitors, with the aim of establishing a foundation for their future inclusion as a standard treatment option. Number of annotations: 7 (limited to key technical terms) 新しく見つかった「XZF-1」という遺伝子のはたらきを調整する因子（転写因子）は，肺がんができたり進んだりする過程に関わっている可能性が，近年の実験中心の研究で示されてきた．（中略）こうした結果をもとに，XZF-1 阻害剤の効果を評価する試験計画を整え，将来的に標準的な治療選択肢として導入するための基盤づくりを進めている．
Gemini	Recent basic research has shown that XZF-1, a newly discovered transcription factor, plays an important role in how lung cancer begins and progresses. (omitted) In addition, preparations are underway to create a detailed protocol for early-stage clinical trials to evaluate the efficacy of XZF-1 inhibitors, with the goal of introducing them in the future as a treatment that is currently established as the most effective and safe (standard treatment). Number of annotations: 22 (including common terms such as “lung cancer” and “efficacy”) 新たに発見された転写因子 XZF-1 が，肺がんの発生や進行において重要な役割を果たすことが基礎研究により示された．（中略）さらに，XZF-1 阻害剤の有効性を評価する初期臨床試験の詳細な計画（プロトコール）を策定し，将来的に標準治療として導入するための準備を進めている．

Table 4: Comparison of generation outputs between GPT-5.2 and Gemini under the Original+GLFS setting. GPT-5.2 produces 7 annotations limited to key technical terms, whereas Gemini generates approximately three times more annotations (22), including common terms such as “lung cancer” and “efficacy.”

6. Results and Discussion

Table 3 presents the automatic evaluation results. Prompt design had a notable effect on the scores. The single-stage setting (**Original+GLFS**) outperformed the two-stage setting (**Simp+GLFS**) across most metrics. GPT-5.2 achieved the highest BLEU scores, while DeepSeek and Gemini showed greater variation across settings.

6.1. Single-stage vs. Two-stage Approaches

The two-stage approach yielded lower performance. The Base stage simplifies technical terms; for instance, “EGFR mutation” becomes “specific gene mutations” (Table 3). Once simplified, these terms are difficult to recover in the Refinement stage, leading to information loss.

In contrast, the single-stage approach retains original terminology while adding explanations simultaneously. This preserves both accuracy and readability. These results indicate that direct generation is more effective than stepwise processing for specialized content.

6.2. Model Comparison

GPT-5.2 achieved the highest BLEU scores by making minimal structural changes and limiting annotations to essential terms, consistent with guideline item 4. Gemini, however, produced approximately three times more annotations, including explanations for common terms such as “lung cancer” (Table 3). This verbosity reduced BLEU scores due to lower precision, though ROUGE scores remained stable as the core content was preserved. Controlling annotation density according to reader characteristics remains a key challenge.

6.3. Effect of Few-shot Examples

Comparing Original+GL with Original+GLFS reveals the contribution of few-shot examples to rewriting quality. BLEU scores improved substantially when few-shot examples were added: DeepSeek showed a 75% increase (11.1 to 19.6), Gemini 52% (8.9 to 13.5), and GPT-5.2 6% (12.3 to 13.1). BERTScore also improved consistently across all models, indicating enhanced semantic alignment with reference texts.

These results suggest that guidelines alone may

be insufficient for LLMs to infer the expected output format and annotation style. Few-shot examples provide concrete demonstrations of how guidelines should be applied, reducing ambiguity in interpretation. The improvement was most pronounced for DeepSeek, suggesting that models with stronger in-context learning capabilities may benefit more from example-based guidance.

Interestingly, Gemini exhibited a decrease in ROUGE scores under GLFS compared to GL alone (ROUGE-1: 0.78 to 0.48), despite improvements in BLEU and BERTScore. This discrepancy may reflect Gemini's tendency to generate additional explanatory content when provided with examples, which increases output length and reduces n-gram overlap with the reference.

These findings also point to directions for future research. Constructing a richer reference dataset authored by multiple writers would enable a more objective evaluation that does not rely on the writing style or judgment of a single individual. Furthermore, human evaluations involving actual patients and general readers are essential to assess guideline adherence and confirm the absence of medical misunderstandings.

7. Conclusion

Rewriting for non-experts was defined herein as a framework for presenting medical research texts in a form that is accessible to non-expert readers, and its feasibility was investigated using LLMs. The results suggest that directly generating the final output through editing guidelines and concrete examples is effective for achieving rewriting for non-experts.

The goal of the concept targeted in this work is not only to facilitate communication among experts but also to support the creation of a society in which people from diverse backgrounds can access and understand scientific and technological information, including medical knowledge, in a clear and inclusive manner. Future work includes constructing larger datasets covering a broader range of disease domains and exploring mechanisms to ensure the accuracy of medical information.

8. Limitations

This study has several limitations that should be acknowledged. First, the reference texts used for evaluation were created by a single medical writer. There is no single correct answer for rewriting for non-experts; the optimal level of detail and phrasing may vary depending on the assumed knowledge level of the reader and the reading context. Therefore, the current evaluation, based on the degree of alignment with a single writer's output, captures only one dimension of performance.

Second, limitations are inherent in the automatic evaluation metrics used. This study employed n-gram-based metrics such as BLEU and ROUGE, which assess surface-level similarity to reference texts but do not directly measure understandability or adherence to guidelines. For example, in our experiments, Gemini received lower scores due to its tendency to add excessive annotations, which deviated from the guidelines. This outcome is consistent with the behavior of these metrics in penalizing verbose or off-target output. However, even if such annotations were beneficial to readers, they would still be penalized unless included in the reference. Thus, current metrics make it difficult to distinguish between unnecessary redundancy and useful supplementary information.

Regarding the reviewer's question of whether non-expert human evaluation is necessary, we consider it essential for this task. Because rewriting for non-experts targets perceived clarity, trust, and actionable understanding, automatic overlap metrics alone are insufficient to establish practical usefulness. In this study, we prioritized controlled comparisons across prompting settings as an initial step; however, a follow-up evaluation with non-expert participants is required to assess comprehensibility, perceived safety, and guideline appropriateness. We plan to include such human evaluation as a core component of the next experimental phase.

9. Acknowledgments

This study was supported by AMED Grant Number JP25oa0439009, the Strategic Innovation Promotion Program (SIP) "Development of an Integrated Healthcare System" JPJ012425, and JSPS Research Start-up Support JP25K24412.

10. Bibliographical References

Sweta Agrawal and Marine Carpuat. 2024. *Transactions of the Association for Computational Linguistics*, 12:432–448.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. The (un)suitability of automatic evaluation metrics for text simplification. *Computational Linguistics*, 47(4):861–889.

Scott Crossley, Aron Heintz, Joon Suh Choi, Jordan Batchelor, Mehrnoush Karimi, and Agnes Malatinszky. 2022. A large-scaled corpus for

assessing text readability. *Behavior Research Methods*, 55:1023–1047.

DeepSeek-AI. 2025. Deepseek-v3.2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.

Gemini Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. Technical report, Google. Accessed: 2026-01-08.

Gerd Gigerenzer and Adrian Edwards. 2003. Simple tools for understanding risks: From innumeracy to insight. *BMJ: British Medical Journal*, 327(7417):741–744.

Thomas Goldsack, Seán Cartright, Jon Chamberlain, and Massimo Poesio. 2022. Making science simple: Corpora for the lay summarisation of scientific literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Thomas Goldsack, Jon Chamberlain, and Massimo Poesio. 2024. Overview of the BioLaySumm 2024 shared task on the lay summarization of biomedical research articles. In *Proceedings of the 2024 BioNLP Workshop*, Bangkok, Thailand. Association for Computational Linguistics.

Rachael Gotlieb, Corinne Praska, Marissa A. Hendrickson, Jordan Marmet, Victoria Charpentier, Emily Hause, Katherine A. Allen, Scott Lunos, and Michael B. Pitt. 2022. Accuracy in patient understanding of common medical phrases. *JAMA Network Open*.

Yue Guo, Wei Qiu, Yizhong Wang, and Trevor Cohen. 2021. Automated lay language summarization of biomedical scientific reviews. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(1):160–168.

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. Research Branch Report 8-75, Chief of Naval Technical Training, Naval Air Station Memphis, Memphis, TN.

Gondy Leroy, David Kauchak, and Obay Mouradi. 2013. A user-study measuring the effects of lexical simplification and coherence enhancement on perceived and actual text difficulty. *International Journal of Medical Informatics*, 82(8):717–730.

Qing Lu and Peter J. Schulz. 2024. Physician perspectives on internet-informed patients: Systematic review. *Journal of Medical Internet Research*, 26(1):e47620.

Ziqi Luo, Shervin Radpour, Shihua Wang, Sarah Yeston, Yining Ruan, and Mark Gerstein. 2024. The lay person's guide to biomedicine: Orchestrating large language models. *arXiv preprint arXiv:2402.13498*.

Louis Martin, Éric de la Clergerie, Benoît Sagot, and Antoine Bordes. 2020. Controllable sentence simplification. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4689–4698, Marseille, France. European Language Resources Association.

Brian Ondov, Kush Attal, and Dina Demner-Fushman. 2022. A survey of automated methods for biomedical text simplification. *Journal of the American Medical Informatics Association: JAMIA*, 29(11):1976–1988.

OpenAI. 2025. Introducing gpt-5.2. Accessed: 2026-01-08.

Jessica Root, Nancy V. Oster, Shanda L. Jackson, Ralitza Mejilla, Jan Walker, and Joann G. Elmore. 2016. Characteristics of patients who report confusion after reading their primary care clinic notes online. *Health Communication*, 31(6):778–781.

Rahul C. Salvi, Tanvir Alam, Sambuddha Ghosh, Hrishikesh Hegde, Soham Ghosh, Alexis Palmer, and Danielle Mowery. 2025. Towards understanding LLM-generated biomedical lay summaries. In *Proceedings of the 2025 Workshop on Clinical Natural Language Processing for Health (CL4Health)*, Mexico City, Mexico. Association for Computational Linguistics.

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, Melbourne, Australia. Association for Computational Linguistics.

Birte Schmitz. 2023. Improving accessibility of scientific research by artificial intelligence—an example for lay abstract generation. *Digital Health*, 9:20552076231186245.

Gillian Tett. 2015. *The Silo Effect: The Peril of Expertise and the Promise of Breaking Down Barriers*. Simon & Schuster, New York.

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.