Enhancing Accessibility of Medical Texts through Large Language Model-Driven Plain Language Adaptation

Ting-Wei Chang[1] **Hen-Hsen Huang**[2] **Hsin-Hsi Chen**[1,3]

[1]Department of Computer Science and Information Engineering, National Taiwan University, Taiwan
[2]Institute of Information Science, Academia Sinica, Taiwan
[3]AI Research Center (AINTU), National Taiwan University, Taiwan

changtw@nlg.csie.ntu.edu.tw hhhuang@iis.sinica.edu.tw hhchen@ntu.edu.tw

Abstract

This paper addresses the challenge of making complex healthcare information more accessible through automated Plain Language Adaptation (PLA). PLA aims to simplify technical medical language, bridging a critical gap between the complexity of healthcare texts and patients' reading comprehension. Recent advances in Large Language Models (LLMs), such as GPT and BART, have opened new possibilities for PLA, especially in zero-shot and few-shot learning contexts where task-specific data is limited. In this work, we leverage the capabilities of LLMs such as GPT-4o-mini, Gemini-1.5-pro, and LLaMA for text simplification. Additionally, we incorporate Mixture-of-Agents (MoA) techniques to enhance adaptability and robustness in PLA tasks. Key contributions include a comparative analysis of prompting strategies, finetuning with QLoRA on different LLMs, and the integration of MoA technique. Our findings demonstrate the effectiveness of LLM-driven PLA, showcasing its potential in making healthcare information more comprehensible while preserving essential content.

1 Introduction

Healthcare information is often presented in complex, technical language that can be challenging for the general public to understand. Yet, research highlights a persistent gap between the language used in medical documentation and the reading comprehension abilities of patients (Guo et al., 2020; Goldsack et al., 2022; Luo et al., 2022; Goldsack et al., 2023a). Hence, PLA aims to bridge this gap by simplifying complex healthcare texts, making vital health information more accessible and promoting better health outcomes (Goldsack et al., 2023a; McCray, 2004).

Developing automated solutions for PLA in medical terminology faces considerable challenges due to the complexity of extracting and simplifying specialized terms within biomedical texts. Recent methods have begun to incorporate semantic-sensitive approaches alongside word embedding techniques, such as the neighborhood context-based "Snowball" method, which has demonstrated promise in initial evaluations with expert-validated standards for term extraction (Bay et al., 2020). Another study attempts to provide comprehensive background explanations for key medical concepts in abstracts, relying on LLMs to accurately identify and interpret these concepts (Luo et al., 2024).

Recently, advancements in LLMs, such as T5, GPT and BART, have transformed various Natural Language Processing (NLP) tasks, offering potential solutions to PLA challenges (Li et al., 2024; Knappich et al., 2023). LLMs demonstrate promising abilities in lay summarization, particularly in zero-shot and few-shot learning contexts, where minimal task-specific data is available (Turbitt et al., 2023). Other research addresses this problem by finetuning LLMs to improve their performance in enhancing the readability of biomedical texts (Li et al., 2024; Knappich et al., 2023; Sim et al., 2023; Reddy et al., 2023).

This paper compares and utilizes the latest models, including GPT-4o-mini, Gemini-1.5-pro, LLaMA, Gemma, and Mistral, along with techniques such as zero-shot, few-shot, QLoRA finetuning, and the advanced MoA methodology, which leverages the collective strengths of multiple LLMs (Wang et al., 2024). MoA has demonstrated superior performance across various benchmarks, making it a compelling choice for complex tasks like PLA. The primary contributions of this work are outlined as follows:

1. Comparison of prompting techniques: We conduct a thorough comparison of different prompting strategies, including zero-shot prompting, in-context learning (ICL), and ICL with semantic similarity (Liu et al., 2021).

2. Evaluation of finetuning approaches: We explore the effectiveness of finetuning advanced LLMs using QLoRA (Dettmers et al., 2023), assessing how this technique impacts the models' ability to generate simplified medical language.

3. Integration of MoA techniques: We utilize the MoA technique, effectively combining the unique capabilities of various LLMs to compare and analyze its impact on overall performance in PLA tasks.

4. Automated evaluation and LLM judging: To enhance the robustness of our results, we implement an automated evaluation framework alongside LLM judging mechanisms. This dual approach allows for a comprehensive analysis of the outputs, ensuring that our findings are both reliable and actionable.

5. We evaluate the new term replacement task introduced in PLABA 2024 using various prompt techniques.

2 Related Work

In this section, we first introduce previous Plain Language Adaptation of Biomedical Abstracts (PLABA) work (Attal et al., 2023), then introduce biomedical text simplification, followed by lay summarization, and finally the efficient finetuning approach used in our work.

2.1 PLABA

The PLABA dataset addresses the challenge of simplifying complex biomedical literature in sentence level for general audiences. Despite the availability of health-related resources like MedlinePlus [1], many scientific articles remain inaccessible to the public due to their specialized language. While various efforts have aimed to adapt technical terms for readability, creating manual plain language summaries for every medical article is impractical. Automated adaptation, supported by language models, has emerged as a promising solution, but requires high-quality, sentence-aligned datasets to train effective models. Existing datasets often suffer from imperfect alignments or lack sufficient scale, making them suboptimal for training and evaluation. PLABA fills this gap by providing 750 manually adapted abstracts from PubMed, offering sentence-level alignment for 7,643 pairs, each crafted by expert annotators. This dataset enables document- and sentence-level simplification and incorporates sentence-splitting to improve readability, making it a valuable gold standard for evaluating adaptation techniques in biomedical text simplification.

[1] https://medlineplus.gov/

2.2 Biomedical text simplification

Language models like LLaMA-2 have been effectively applied to this task. For example, a LLaMA 2-based system achieved top performance in the previous PLABA shared task, focusing on simplifying complex biomedical text (Knappich et al., 2023). This approach highlights the difficulty in training models with high token overlap between input and output texts, which can limit the model's ability to perform substantive edits. To address this, sentence- and token-level loss weights were introduced, giving more emphasis to modified tokens, which led to simplifications closely aligned with human-generated adaptations, showing improved SARI and FKGL scores.

Another study explored various powerful LLMs, including encoder-decoder models (T5, SciFive, BART), GPT models (GPT-3.5, GPT-4), and control-token mechanisms within BART-based models, to simplify biomedical abstracts using the PLABA dataset (Li et al., 2024). Through domain-specific finetuning and prompt-based learning, these models were evaluated on both automated and human metrics. The BART-Large model with control token achieved the highest SARI score, while T5-Base scored best on BERTScore, balancing simplicity and meaning preservation.

2.3 Lay summarization

The BioLaySumm shared tasks in 2023 and 2024 represent significant efforts in advancing lay summarization of biomedical research (Goldsack et al., 2023b, 2024). These tasks, unlike PLABA tasks, focus on generating comprehensible summaries for non-expert audiences by using abstractive summarization techniques. In the 2023 task, models were trained to produce "lay summaries" that capture the essence of a full article while remaining accessible to general readers. Building on its success, the 2024 edition expanded participation and saw a trend towards innovative approaches, particularly with LLMs, reflecting the growing emphasis on this area.

2.4 Efficient finetuning

Efficient finetuning techniques like QLoRA are crucial for reducing computation costs and accelerating the finetuning process while retaining the full performance of 16-bit finetuning (Dettmers et al., 2023). This is achieved by using a 4-bit quantized model and Low Rank Adapters (LORA) to backpropagate gradients without unfreezing the pretrained model weights (Hu et al., 2021), leading to a substantial reduction in memory usage without sacrificing output quality. QLoRA's efficiency allows extensive experimentation across multiple model architectures, parameters, and instruction-following datasets, highlighting that high-quality, smaller datasets often outperform large, less-focused datasets in instruction finetuning. QLoRA's open-source release, along with comprehensive analyses, provides a valuable framework for efficient, high-performance finetuning in natural language processing.

2.5 MoA

The MoA methodology represents an innovative approach in leveraging the combined expertise of multiple LLMs to enhance natural language understanding and generation tasks (Wang et al., 2024). Unlike traditional single-model setups, MoA organizes LLMs into layers, with each agent in a given layer receiving input from the outputs of agents in the previous layer. This collaborative structure capitalizes on the "collaborativeness" phenomenon—where LLMs produce improved responses when they can build upon outputs from other models, even those of lower quality. MoA achieves state-of-the-art performance on several benchmarks and surpasses leading models like GPT-4 Omni.

3 Methodologies

The overall framework of our experimental design is illustrated in Figures 1 and 2, which outline the methodologies for the PLA task and the term replacement task.

The PLA Task aims to simplify each abstract while keeping sentences separate and ensuring accuracy and clarity for general readers. To accomplish this, we employ models using zero-shot prompting, random few-shot ICL, few-shot ICL enhanced by semantic similarity, and a zero-shot approach following QLoRA finetuning of LLMs on training data.

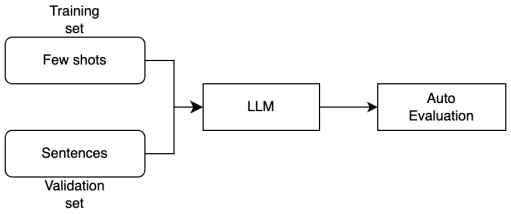

Figure 1: Pipeline of the PLA task in a few-shot setting. The provided annotated data is split into training and validation sets.

New to PLABA 2024, the Term Replacement Task focuses on identifying and simplifying technical terms within abstracts. Our approach starts by identifying complex terms in medical texts, then classifies the type of replacement needed, and finally generates lay language equivalents. This process is performed using few-shot ICL, with a focus on enhancing semantic similarity for more accurate replacements.

3.1 Models

Our experiments utilize a diverse set of advanced LLMs, each with unique capabilities suited to NLP tasks.

3.1.1 Gemini

Gemini-1.0-pro, Gemini-1.5-flash, Gemini-1.5-pro are developed by Google, the Gemini models are multimodal, capable of handling image, audio, video, and text. Gemini-1.5 models, such as Pro and Flash, are particularly known for their efficiency in handling large contexts, enabling detailed recall over long text and multimedia inputs.

3.1.2 GPT-4o-mini

A cost-efficient model from OpenAI, GPT-4o-mini is optimized for high performance on natural language tasks while lowering the costs and latency.

3.1.3 LLaMA

LLaMA2 (7B), LLaMA3 (8B, 70B), LLaMA3.1 (8B, 70B) are developed by Meta, offer enhanced support for multilinguality, reasoning, and long-context processing. LLaMA 3.1 supports up to 128K tokens, allowing for complex understanding and generation tasks.

3.1.4 Gemma 2

The Gemma 2 (2B, 9B, 27B) series from Google DeepMind is a lightweight yet capable family of models, scaling up to 27 billion parameters. These

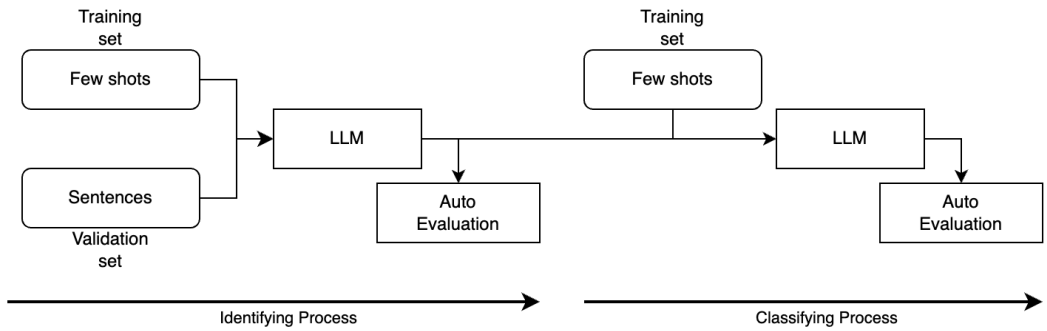

Figure 2: Pipeline of term replacement, containing Identifying process and Classifying process. The provided annotated data is split into training and validation sets.

models leverage advanced techniques such as interleaving local-global attentions and grouped-query attention, along with knowledge distillation, enabling smaller models to achieve performance comparable to larger ones.

3.1.5 Mistral

The Mistral 7B model uses grouped-query attention and sliding window attention, which significantly improves inference speed and allows handling of longer sequences, making it suitable for real-time applications. Mistral NeMo, developed in collaboration with NVIDIA, is a larger model designed to handle even more extensive tasks with a context window of up to 128K tokens.

3.2 QLoRA

In our experiments, we applied QLoRA finetuning to optimize model efficiency and memory usage. The finetuning process employed a 4-bit quantized model to significantly reduce computational load. We set both the LoRA rank and LoRA alpha to 16, used a learning rate of 2e-4, kept the training batch size minimal at 2, and completed the finetuning in a single epoch.

3.3 MoA

We apply the MoA approach across different settings, utilizing Gemini-1.5-Flash to perform Aggregate-and-Synthesize while simplifying the process by employing only a single-layer MoA. Our approach builds upon the 5-shot setting from previous steps, using the following models as the adaptation models: Gemini-1.0-Pro, Gemini-1.5-Flash, Gemini-1.5-Pro, Gemma-2-27B, GPT-4o-Mini, Meta-Llama-3.1-8B, and Mistral-Nemo-Instruct-2407. Additionally, we explore the integration of finetuned models within the MoA framework by incorporating Gemma-2-27B, Meta-Llama-3.1-8B, and Mistral-Nemo-Instruct-2407, allowing us to assess the impact of finetuning on MoA's performance in PLA tasks.

3.4 Metrics for PLA task

We apply automatic evaluation metrics to assess the performance of the PLA task, referencing the scoring methods used in PLABA2023 works and BioLaySumm (Li et al., 2024; Goldsack et al., 2024, 2023b).

- Relevance: BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), BERTScore (Zhang et al., 2020), and SARI (Xu et al., 2016).
- Readability: Flesch-Kincaid Grade Level (FKGL), Dale-Chall Readability Score (DCRS), Coleman-Laiu Index (CLI).
- Factuality: AlignScore (Zha et al., 2023), SummaC (Laban et al., 2021).
- LLM Judge: Simplicity, Accuracy, Completeness, Brevity [2].

Metrics like BLEU, ROUGE, BERTScore, and SARI are used to evaluate how closely the generated text aligns with reference summaries or simplified texts. BLEU and ROUGE measure the overlap of n-grams between generated and reference texts, with BLEU focusing on precision and ROUGE on recall. BERTScore uses contextual embeddings to evaluate semantic similarity, capturing meaning beyond surface-level token matches. SARI specifically assesses simplification by measuring edits made to the source text to match references, rewarding appropriate additions, deletions, and mod-

[2] https://bionlp.nlm.nih.gov/plaba2024/

ifications, making it particularly suited for PLA tasks.

To gauge the accessibility of the text, readability metrics such as the Flesch-Kincaid Grade Level (FKGL), Dale-Chall Readability Score (DCRS), and Coleman-Liau Index (CLI) are employed. FKGL estimates the education level needed to understand the text, DCRS assesses readability based on word familiarity, and CLI evaluates readability through character and sentence length, with all three metrics indicating if the text is accessible to a general audience. In all three metrics, lower scores indicate greater simplicity and ease of comprehension.

For ensuring factual consistency, we apply AlignScore and SummaC. AlignScore is a versatile metric developed to handle factual inconsistency across diverse input-output pairs, leveraging a unified training approach that integrates data from multiple tasks, including NLI and QA, thus enhancing its generalizability. SummaC is designed specifically for summarization, using sentence-level analysis within documents to aggregate consistency scores. Both metrics provide a robust framework for detecting contradictions and ensuring the generated text remains factually aligned with the input information.

For LLM Judge, we utilized Gemini-1.5-flash and referenced prompts from previous work and made modifications to better suit our needs (Luo et al., 2024). The evaluation criteria were refined to focus on four main aspects: Simplicity, Accuracy, Completeness, and Brevity.

3.5 Metrics for term replacement task

After identifying difficult terms, we evaluate performance using the F1 score. For replacement type classification, we use the multilabel F1 score to measure accuracy. As for generating replacements, since no prior work offers a suitable reference, we did not conduct automatic evaluation for this aspect in this study.

4 Experiments and Evaluations

The PLABA dataset includes 750 biomedical abstracts that have been manually adapted into plain language by annotators, totaling 7,643 sentence pairs. In the 2024 PLABA track, additional tasks build upon this foundation. For the PLA task, 40 new consumer questions are introduced, each accompanied by 10 corresponding abstracts, resulting in a total of 400 test cases. In the Term Replacement Task, introduced exclusively in the PLABA 2024 track, 40 questions were selected from the original PLABA dataset. Each question is linked to 10 abstracts, with 10 questions provided as training data and the remaining 30 designated as test cases.

4.1 Data preprocessing and splitting

In the PLA task, we followed the data-splitting methodology outlined in the original PLABA dataset paper (Attal et al., 2023). The 921 abstracts were divided into approximately 85% for the training set and 15% for the validation set. Each topic was grouped and contained exclusively within either the training or test set to ensure unbiased evaluation. For the term replacement task, we split the provided annotated data into a 70/30 ratio.

4.2 PLA task

4.2.1 Baseline from previous work

We implemented the method from the original work (Attal et al., 2023) and applied our evaluation method to establish a baseline comparison. The results are presented in Table 1.

4.2.2 Zero-shot vs ICL

Table 2 compares Gemini-1.0-pro, Gemini-1.5-flash, and GPT-4o-mini across different prompt methods. It shows that using few-shot ICL improves Relevance score, with ICL using semantic similarity performing slightly better than random few-shot selection. Different few-shot methods have minimal impact on the Readability and Factuality metrics.

4.2.3 Finetuning

Table 3 presents a comparison of various model series after finetuning. Results indicate a comprehensive improvement in Relevance and Factuality scores across all models. While Readability scores also increased. Using Gemini-1.5-flash as the LLM Judge, we observe that Simplicity remains relatively consistent, while Accuracy generally improves. Completeness and Brevity metrics show a uniform increase across models, highlighting the effectiveness of finetuning in enhancing these aspects.

4.2.4 MoA

Table 4 presents a comparison of various model series after finetuning. Results indicate a comprehensive improvement in Relevance and Factual-

Model	BLEU	SARI	R-1	R-2	R-L	BertS	FKGL	DCRS	CLI	AlignS	SummaC
T0PP	0.90	32.97	24.07	12.60	22.35	85.55	14.68	13.29	16.19	68.86	44.01
Bart-base	24.37	40.13	59.39	30.41	56.71	89.21	12.89	**10.88**	13.85	61.53	45.71
Bart-large-cnn	24.31	**40.29**	59.32	31.09	56.72	89.25	12.72	11.20	**13.83**	70.42	50.57
Bart-large	24.62	40.27	59.56	31.03	56.94	89.42	**12.60**	10.99	13.88	69.39	50.83
Pegasus-large	**25.83**	37.70	**60.47**	**32.51**	**57.88**	**89.70**	12.93	11.00	14.40	**74.80**	**55.89**

Table 1: Baseline evaluation scores for the models mentioned in the original paper. T0PP is not finetuned, while the others are finetuned models. In FKGL,DCRS,CLI, lower scores indicate greater simplicity and ease of comprehension. R = ROUGE, BertS = BertScore, AlignS =AlignScore

Model	Type	FewShot	BLEU	SARI	R-1	R-2	R-L	BertS	FKGL	DCRS	CLI	AlignS	SummaC
Gemini-1.0-pro	Zero	0	16.27	42.35	48.37	23.70	41.57	91.57	9.36	11.36	11.44	82.45	49.66
Gemini-1.0-pro	Closest	1	23.73	42.59	54.63	32.25	49.23	92.35	11.27	12.01	12.86	85.13	57.91
Gemini-1.0-pro	Closest	2	24.81	42.51	55.33	33.24	50.04	92.43	11.43	11.93	12.76	**86.01**	58.85
Gemini-1.0-pro	Closest	3	26.23	42.52	56.41	34.45	51.37	92.57	11.67	12.08	12.94	85.99	**60.26**
Gemini-1.0-pro	Closest	4	26.62	41.02	56.57	34.91	51.53	92.53	11.80	12.15	13.10	85.85	59.86
Gemini-1.0-pro	Closest	5	**27.26**	41.45	**57.08**	**35.53**	**52.19**	**92.59**	11.99	12.17	13.21	85.50	60.08
Gemini-1.5-flash	Zero	0	14.15	43.21	46.82	21.47	40.58	91.42	**9.97**	**10.88**	**11.16**	71.05	36.92
Gemini-1.5-flash	Random	1	16.90	44.00	49.03	24.12	42.83	91.72	10.39	11.21	11.60	74.76	41.01
Gemini-1.5-flash	Random	2	18.17	44.11	50.19	25.30	44.10	91.84	10.34	11.29	11.78	74.47	41.16
Gemini-1.5-flash	Random	3	19.31	45.07	51.51	26.72	45.42	92.02	10.49	11.36	11.84	75.18	42.94
Gemini-1.5-flash	Random	4	20.18	44.77	52.34	28.15	46.68	92.06	10.84	11.43	11.98	74.13	43.09
Gemini-1.5-flash	Random	5	21.24	45.51	53.06	28.97	47.36	92.13	11.05	11.52	12.08	74.29	43.71
Gemini-1.5-flash	Closest	1	16.97	44.21	49.11	24.33	42.86	91.70	10.31	11.10	11.49	73.99	40.87
Gemini-1.5-flash	Closest	2	18.85	45.30	50.55	26.12	44.41	91.82	10.44	11.19	11.57	74.08	41.78
Gemini-1.5-flash	Closest	3	19.81	45.59	51.77	27.47	45.88	92.00	10.67	11.25	11.72	72.51	41.88
Gemini-1.5-flash	Closest	4	21.00	**46.11**	52.98	28.95	47.07	92.08	10.93	11.37	11.93	73.09	42.75
Gemini-1.5-flash	Closest	5	21.37	46.04	53.57	29.70	47.97	92.08	11.11	11.43	11.98	71.91	42.47
GPT-4o-mini	Zero	0	12.37	41.95	45.07	18.89	38.00	91.31	11.22	11.30	11.52	76.24	38.09
GPT-4o-mini	Random	1	15.58	43.65	48.02	22.20	41.15	91.76	11.20	11.57	11.96	79.45	41.86
GPT-4o-mini	Random	2	16.62	44.03	49.02	23.21	42.48	91.90	11.43	11.74	12.19	79.89	42.53
GPT-4o-mini	Random	3	17.05	44.41	49.54	23.73	43.00	91.97	11.51	11.77	12.16	81.01	43.39
GPT-4o-mini	Random	4	17.36	44.67	50.33	24.19	43.45	92.00	11.64	11.81	12.21	80.66	43.73
GPT-4o-mini	Random	5	18.15	45.07	50.85	25.01	44.10	92.05	11.79	11.83	12.33	80.47	43.78
GPT-4o-mini	Closest	1	16.10	44.17	48.57	22.67	41.77	91.82	11.43	11.61	11.91	79.63	42.19
GPT-4o-mini	Closest	2	16.85	44.72	49.48	23.68	42.89	91.90	11.53	11.67	11.92	79.21	42.03
GPT-4o-mini	Closest	3	17.45	45.07	50.34	24.40	43.74	91.99	11.51	11.67	11.99	80.11	42.55
GPT-4o-mini	Closest	4	18.08	45.47	50.97	25.11	44.47	92.08	11.60	11.72	12.05	79.91	43.14
GPT-4o-mini	Closest	5	18.44	45.62	51.34	25.70	44.91	92.12	11.67	11.74	12.11	79.62	42.95

Table 2: Auto Evaluation of zero-shot versus different few shots type. "Type" refers to the few-shot ICL method: "Zero" indicates zero-shot, "Closest" represents few-shot ICL with closest semantic similarity, and "Random" indicates randomly selected few shots. "FewShot" indicates the number of few shots. In FKGL,DCRS,CLI, lower scores indicate greater simplicity and ease of comprehension. R = ROUGE, BertS = BertScore, AlignS = AlignScore

ity scores across all models. While Readability scores also increased. Using Gemini-1.5-flash as the LLM Judge, we observe that Simplicity remains relatively consistent, while Accuracy generally improves. Completeness and Brevity metrics show a uniform increase across models, highlighting the effectiveness of finetuning in enhancing these aspects.

4.3 Term replacement task

Table 5 shows the evaluation of term replacement task. Compared to zero-shot, using few-shot ICL with the closest semantic similarity improves performance in both identifying difficult terms and classifying replacement types.

4.4 Shared Task Results on Test Dataset

Table 6 and Figure 3 present the results of the test dataset as evaluated by the organizers (Ondov et al., 2024) for the PLA and term replacement tasks. The evaluation was conducted through human assessment, measuring four key criteria: Accuracy, Completeness, Simplicity, and Brevity, each scored on a scale of -1, 0, or 1. The final scores represent the average ratings across all evaluated samples. Some values remain unavailable due to the organizers' incomplete evaluation, attributed to a limited number of judges.

Model	FT	BLEU	SARI	R-1	R-2	R-L	BertS	FKGL	DCRS	CLI	AlignS	SummaC	Sim	Acc	Com	Bre
Llama-2-7b-chat	X	7.36	40.58	35.92	14.77	30.79	88.92	10.62	10.36	11.63	46.00	31.68	8.22	8.46	8.09	6.32
Llama-2-7b-chat	O	24.79	39.42	51.67	32.87	48.04	91.21	12.16	12.83	15.16	58.20	48.32	8.47	8.32	8.13	8.21
Llama-3-8b-Instruct	X	11.44	41.33	44.88	18.62	36.80	91.01	11.20	11.28	12.11	73.58	38.50	8.83	9.06	8.69	7.88
Llama-3-8b-Instruct	O	33.77	35.31	61.67	**43.04**	58.44	93.14	13.17	12.79	14.47	88.82	70.85	8.70	9.46	9.48	9.12
Llama-3-70b-Instruct	X	10.79	40.77	42.24	16.45	35.02	90.90	10.82	10.77	10.94	72.52	34.22	8.86	9.01	8.60	7.92
Llama-3-70b-Instruct	O	33.38	37.84	61.81	42.82	58.48	**93.23**	13.00	12.67	14.29	86.63	67.84	8.77	9.38	9.37	**9.13**
Llama-3.1-8B-Instruct	X	9.41	38.96	39.69	14.80	32.82	90.41	9.76	10.53	10.87	71.03	34.43	8.84	8.59	7.99	8.03
Llama-3.1-8B-Instruct	O	31.75	40.57	60.19	40.21	56.40	92.95	12.50	12.41	13.75	80.79	61.38	8.79	9.24	9.15	8.82
Llama-3.1-70B-Instruct	X	13.25	42.81	45.05	19.97	38.18	91.18	10.26	10.52	10.40	73.52	38.17	8.82	9.03	8.63	7.95
Llama-3.1-70B-Instruct	O	31.85	41.39	60.87	41.17	57.14	93.07	12.51	12.41	13.80	81.61	61.12	8.81	9.20	9.09	8.90
Mistral-7b-instruct-v0.3	X	15.08	41.81	47.45	21.87	39.83	91.41	10.60	11.69	12.23	84.74	48.31	8.91	9.18	8.88	8.30
Mistral-7b-instruct-v0.3	O	**33.82**	33.75	60.71	42.27	57.61	92.88	13.22	12.78	14.52	87.08	72.15	8.44	9.44	9.42	8.83
Mistral-Nemo-Instruct-2407	X	13.09	41.09	44.36	19.55	38.14	91.35	8.94	11.41	11.34	81.14	44.52	**8.92**	9.10	8.66	8.29
Mistral-Nemo-Instruct-2407	O	31.83	**42.83**	60.37	40.75	56.85	92.98	12.17	12.23	13.48	75.02	56.36	8.90	9.04	8.91	8.84
Gemma-2-2b-it	X	1.97	40.30	34.71	13.04	29.13	89.05	**8.59**	**8.73**	10.36	60.25	31.08	8.18	8.59	8.31	6.39
Gemma-2-2b-it	O	33.03	34.17	61.09	42.25	57.85	93.07	13.06	12.77	14.44	**90.46**	**73.02**	8.60	9.45	9.48	9.08
Gemma-2-9b-it	X	8.73	38.89	38.73	14.14	32.07	90.32	9.17	10.09	**10.15**	65.73	31.64	8.80	8.93	8.46	7.67
Gemma-2-9b-it	O	33.51	35.35	61.66	42.82	58.41	93.12	13.09	12.76	14.46	88.52	70.92	8.68	**9.47**	**9.48**	9.08
Gemma-2-27b-it	X	10.29	39.75	41.41	15.75	34.19	90.76	8.89	10.59	10.52	71.31	34.35	8.87	9.05	8.59	7.95
Gemma-2-27b-it	O	33.78	36.80	**61.84**	42.90	**58.60**	93.20	12.99	12.66	14.34	86.15	68.68	8.66	9.46	9.40	9.00

Table 3: Auto Evaluation of LLMs w/ or w/o finetuning. "FT" indicates whether finetuning is applied. R = ROUGE, BertS = BertScore, AlignS =AlignScore, Sim = Simplicity, Acc = Accuracy, Com = Completeness, Bre = Brevity

Model	FT	BLEU	SARI	R-1	R-2	R-L	BertS	FKGL	DCRS	CLI	AlignS	SummaC
MoA	X	8.49	38.35	40.43	15.05	33.0	90.56	9.07	12.16	12.74	53.88	32.64
MoA	O	8.21	38.05	40.57	15.14	33.99	90.53	9.08	12.27	12.83	54.84	32.52

Table 4: Auto Evaluation of MoA w/ or w/o finetuned models. "FT" indicates whether finetund models are included. R = ROUGE, BertS = BertScore, AlignS =AlignScore

Model	Identify shots	Idendify (F1)	Classify shots	Classify (F1)
Gemini-1.5-flash	0	27.36	0	53.87
Gemini-1.5-flash	0	27.36	1	52.93
Gemini-1.5-flash	0	27.36	3	59.22
Gemini-1.5-flash	0	27.36	5	65.63
Gemini-1.5-flash	1	35.99	0	54.38
Gemini-1.5-flash	1	35.99	**1**	54.45
Gemini-1.5-flash	1	35.99	3	61.22
Gemini-1.5-flash	1	35.99	5	66.75
Gemini-1.5-flash	3	35.69	0	56.77
Gemini-1.5-flash	3	35.69	1	56.19
Gemini-1.5-flash	3	35.69	3	63.75
Gemini-1.5-flash	3	35.69	5	**68.52**
Gemini-1.5-flash	5	38.98	0	55.98
Gemini-1.5-flash	5	38.98	1	55.46
Gemini-1.5-flash	5	38.98	3	62.33
Gemini-1.5-flash	5	38.98	5	67.12
Gemini-1.5-pro	0	36.62	0	61.89
Gemini-1.5-pro	0	36.62	1	60.99
Gemini-1.5-pro	0	36.62	3	61.78
Gemini-1.5-pro	0	36.62	5	64.86
Gemini-1.5-pro	1	39.02	0	63.79
Gemini-1.5-pro	1	39.02	1	62.22
Gemini-1.5-pro	1	39.02	3	61.24
Gemini-1.5-pro	1	39.02	5	64.90
Gemini-1.5-pro	3	40.78	0	61.52
Gemini-1.5-pro	3	40.78	1	59.00
Gemini-1.5-pro	3	40.78	3	60.91
Gemini-1.5-pro	3	40.78	5	63.99
Gemini-1.5-pro	5	**41.67**	0	63.31
Gemini-1.5-pro	5	**41.67**	1	62.30
Gemini-1.5-pro	5	**41.67**	3	63.24
Gemini-1.5-pro	5	**41.67**	5	66.96

Table 5: F1 Scores for the Identification process and Multilabel F1 Scores for the Classification process in the Term Replacement Task. All few-shot examples were selected using ICL with closest semantic similarity

For the PLA task, we applied the MoA approach as described earlier. Our method ranked 4th when using finetuned models and 8th without finetuning.

In the term replacement task, all three models employed a 5-shot In-Context Learning (ICL) approach, selecting examples with the highest semantic similarity for both identification and classification. Our team ntu_nlp achieved 2nd place in both the identification task and the overall average score for text generation, demonstrating its effectiveness in structured term adaptation.

5 Discussion

5.1 Readability

From the results in this study, models that were not finetuned showed greater variation in readability scores, likely due to differences in how each model interpreted the prompts. Most of these models tended to generate overly simplified outputs, potentially sacrificing content coherence and detail. Once finetuned, the readability of outputs across different models became more consistent, indicating that finetuning helped the models internalize the style and structure of expert-written plain language texts. This consistency suggests that finetuning enabled the models to adopt expert practices in plain language adaptation, improving the quality

Model	Acc	Com	Sim	Bre	Avg
GPT	0.9307	0.8118	0.9232	0.8755	0.8853
LLaMA-8B-4bit-MedicalAbstract-seq-to-seq-v1	0.8831	0.8447	0.7792	0.5240	0.7578
LLaMa_3.1_70B_instruction_2nd_run	0.7440	0.7725	0.7044	0.5903	0.7028
TREC2024_SIB_run1	0.9374	0.8614	0.8363	0.6594	0.8236
TREC2024_SIB_run3	0.8170	0.7324	0.8756	0.7392	0.7910
TREC2024_SIB_run4	0.9433	0.8891	0.6965	0.6210	0.7875
UAms-BART-Cochrane	0.9772	0.9466	0.4881	0.6906	0.7756
UAms-ConBART-Cochrane	0.9546	0.9238	0.5653	0.6781	0.7804
bart_base_ft	0.8424	0.7132	0.4500	0.4177	0.6058
gpt-final	0.9537	0.8996	0.7850	0.3782	0.7541
gpt35_dspy	0.9117	0.8700	0.7654	0.6533	0.8001
mistral-FINAL	0.8788	0.8567	0.7188	0.4122	0.7166
mistral-fix	0.8818	0.8603	0.7118	0.4155	0.7174
plaba_um_fhs_sub1	0.8985	0.8427	0.8421	0.7618	0.8363
plaba_um_fhs_sub2	0.9447	0.8588	0.8909	0.7930	0.8718
plaba_um_fhs_sub3	0.9504	0.8681	0.7949	0.6728	0.8215
task2_moa_tier1_post (Ours MoA)	0.8948	0.8463	0.8197	0.6272	0.7970
task2_moa_tier2_post (Ours MoA w/ FT)	0.9307	0.8537	0.8642	0.6468	0.8238

Table 6: PLA Task Results (Ondov et al., 2024): Human evaluations for text generation. "FT" indicates whether finetund models are included. Evaluation metrics include Simplicity (Sim), Accuracy (Acc), Completeness (Com), Brevity (Bre), and an overall Average (Avg).

and readability of the outputs.

5.2 Auto Evaluation

Automated evaluation posed several challenges in this study. One of the primary issues was that standard readability metrics such as FKGL and DCRS are not designed to capture the nuances of plain language adaptation in medical texts. While these metrics provide a general measure of readability, they often fail to reflect the balance between simplifying content and maintaining its accuracy, which is critical in healthcare communication.

5.3 Other Dataset

This study was limited to using the PLABA dataset for training and testing, which may not fully represent the range of medical and healthcare texts encountered in real-world applications. To more comprehensively evaluate the models' capabilities in text simplification, additional datasets specific to other medical domains or containing varied text structures would be beneficial. Access to a broader range of datasets could enable a more robust assessment of the model's ability to generalize across different contexts and further validate its adaptability to varied plain language adaptation tasks in healthcare.

6 Conclusion

In conclusion, this study demonstrates the potential of LLMs to simplify complex medical texts through PLA, leveraging few-shot ICL, finetuning, or the MoA approach. The few-shot approach with the closest semantic similarity improved the models' ability to generate relevant and readable outputs. finetuning also enhanced the effectiveness of PLA, enabling the models to adopt expert writing styles in plain language adaptation. While our results indicate that general-purpose models can effectively adapt to PLA tasks with suitable finetuning, the study also highlights limitations. All models tested were general-purpose LLMs, and none were pre-trained on medical datasets, which may have limited their ability to understand and perform on domain-specific tasks. Future work could explore the impact of models pre-trained on healthcare-specific data to further advance the effectiveness and accuracy of PLA in medical contexts.

					Task 1C (manual judgments, 0-1)				
			Task 1A (F1)	Task 1B (F1)	Accy	Comp	Simp	Brev	Avg.
BU	MLPClassifier-identify-classify-replace-v1	1	0.0459	0.7788	0.7500	0.7258	0.9274	0.9435	0.8367
CLAC	mistral	1	0.441	0.6661	0.9505	0.9483	0.9738	0.6702	0.8857
CLAC	gpt	2	0.3767	0.3795					
IITH	First	1	0.1956	0.7014					
UM	Roberta-base	1	0.4787	0.7765					
ntu_nlp	gemini-1.5-pro_demon5_replace-demon5	1	0.4885	0.6335	0.9117	0.9010	0.9511	0.5871	0.8377
ntu_nlp	gemini-1.5-flash_demon5_replace-demon5	2	0.4431	0.6544					
ntu_nlp	gpt-4o-mini__demon5_replace-demon5	3	0.4518	0.6197					
Yseop	roberta-gbc	1	0.5036	0.6838					

Figure 3: Term Replacement Task Results (Ondov et al., 2024): F1 scores for identifying difficult terms and classifying replacement types, with human evaluations for text generation. Evaluation metrics include Simplicity (Simp), Accuracy (Accy), Completeness (Comp), Brevity (Brev), and an overall Average (Avg.).

References

Kush Attal, Brian Ondov, and Dina Demner-Fushman. 2023. A dataset for plain language adaptation of biomedical abstracts. *Scientific Data*, 10(1):8.

Matthias Bay, Daniel Bruneß, Miriam Herold, Christian Schulze, Michael Guckert, and Mirjam Minor. 2020. Term extraction from medical documents using word embeddings. In *2020 6th IEEE Congress on Information Science and Technology (CiSt)*, pages 328–333.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms.

Tomas Goldsack, Zheheng Luo, Qianqian Xie, Carolina Scarton, Matthew Shardlow, Sophia Ananiadou, and Chenghua Lin. 2023a. Biolaysumm 2023 shared task: Lay summarisation of biomedical research articles. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, page 468–477. Association for Computational Linguistics.

Tomas Goldsack, Zheheng Luo, Qianqian Xie, Carolina Scarton, Matthew Shardlow, Sophia Ananiadou, and Chenghua Lin. 2023b. Overview of the biolaysumm 2023 shared task on lay summarization of biomedical research articles. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 468–477, Toronto, Canada. Association for Computational Linguistics.

Tomas Goldsack, Carolina Scarton, Matthew Shardlow, and Chenghua Lin. 2024. Overview of the BioLaySumm 2024 shared task on the lay summarization of biomedical research articles. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 122–131, Bangkok, Thailand. Association for Computational Linguistics.

Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 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*, pages 10589–10604, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yue Guo, Wei Qiu, Yizhong Wang, and Trevor Cohen. 2020. Automated lay language summarization of biomedical scientific reviews. *CoRR*, abs/2012.12573.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Valentin Knappich, Simon Razniewski, and Annemarie Friedrich. 2023. Boschai @ plaba 2023: Leveraging edit operations in end-to-end neural sentence simplification.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2021. Summac: Re-visiting nli-based models for inconsistency detection in summarization.

Zihao Li, Samuel Belkadi, Nicolo Micheletti, Lifeng Han, Matthew Shardlow, and Goran Nenadic. 2024. Investigating large language models and control mechanisms to improve text readability of biomedical abstracts.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3?

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2022. Readability controllable biomedical document summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4667–4680, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2024. The lay person's guide to biomedicine: Orchestrating large language models.

A. T. McCray. 2004. Promoting Health Literacy. *Journal of the American Medical Informatics Association*, 12(2):152–163.

Brian Ondov, Bill Xia, Ishita Unde, Hoa T. Dang, and Dina Demner-Fushman. 2024. Overview of the TREC 2024 PLABA track. In *The Thirty-Third Text REtrieval Conference Proceedings (TREC 2024)*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*.

Venkat praneeth Reddy, Pinnapu Reddy Harshavardhan Reddy, Karanam Sai Sumedh, and Raksha Sharma. 2023. IITR at BioLaySumm task 1:lay summarization of BioMedical articles using transformers. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 625–628, Toronto, Canada. Association for Computational Linguistics.

Mong Yuan Sim, Xiang Dai, Maciej Rybinski, and Sarvnaz Karimi. 2023. CSIRO Data61 team at BioLaySumm task 1: Lay summarisation of biomedical research articles using generative models. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 629–635, Toronto, Canada. Association for Computational Linguistics.

Oisín Turbitt, Robert Bevan, and Mouhamad Aboshokor. 2023. MDC at BioLaySumm task 1: Evaluating GPT models for biomedical lay summarization. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 611–619, Toronto, Canada. Association for Computational Linguistics.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. 2024. Mixture-of-agents enhances large language model capabilities.

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.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.