LaySummX at BioLaySumm: Retrieval-Augmented Fine-Tuning for Biomedical Lay Summarization Using Abstracts and Retrieved Full-Text Context

Fan Lin[*] **and Dezhi Yu**[*]
School of Information, University of California, Berkeley, USA
{fan.lin, dezhi.yu}@berkeley.edu

Abstract

Generating lay summaries of biomedical research remains a time-intensive task, despite their importance in bridging the gap between scientific findings and non-expert audiences. This study introduces a retrieval-augmented fine-tuning framework for biomedical lay summarization, integrating abstract-driven semantic retrieval with LoRA-tuned LLaMA 3.1 models. Abstracts are used as queries to retrieve relevant text segments from full-text articles, which are then incorporated into prompts for supervised fine-tuning. Evaluations on the PLOS and eLife datasets show that this hybrid approach significantly improves relevance and factuality metrics compared to both base models and those tuned individually, while maintaining competitive readability. Prompt design experiments highlight a trade-off between readability and factual accuracy. Our fine-tuned model demonstrates strong performance in relevance and factuality among open-source systems and rivals closed-source models such as GPT, providing an efficient and effective solution for domain-specific lay summarization.

1 Introduction

Biomedical research is essential to advancing human health and societal well-being. However, with over 1.5 million articles published annually (González-Márquez et al., 2024), it is increasingly difficult for readers to absorb new findings efficiently. Although abstracts are designed to summarize key results, their technical language often limits accessibility for non-experts. Lay summaries help bridge this gap by presenting core contributions in clear, non-technical language, yet they remain uncommon due to the manual effort required. The BioLaySumm shared task addresses this challenge by promoting the automatic generation of high-quality lay summaries to support broader understanding of biomedical research (Xiao et al., 2025).

Recent advances in large language models (LLMs) have enabled zero- and few-shot summarization, reshaping the field through strong language understanding and instruction-following capabilities (Zhang et al., 2024). Results from the BioLaySumm shared task further demonstrate that LLM-based methods perform well in generating lay summaries of biomedical texts (Goldsack et al., 2024, 2023).

One of the key challenges in the BioLaySumm shared task is the computational cost of feeding an entire research article into a large language model (LLM), even though many recent LLMs support extended context windows (e.g., up to 128k tokens in LLaMA 3.1). Prior research has investigated several strategies to address this issue, including text chunking, which segments lengthy documents into smaller, more manageable units for summarization by models such as Mixtral 8x7B (Bao et al., 2024), or extractive summarization techniques that identify and select salient sentences from the full text (You et al., 2024).

In this study, we developed a workflow that integrates retrieval-augmented generation (RAG) with LoRA-based fine-tuning to improve the performance of LLaMA 3.1 on the biomedical lay summarization task (Figure 1) [1]. To address input length constraints imposed by limited GPU memory, we used the abstract of each article as a query to retrieve relevant but complementary content from the full text. Both the abstract and the retrieved information were used to fine-tune the model, enabling it to generate lay summaries that match the editorial style of the target journals, PLOS and eLife.

[*] These authors contributed equally.

[1] https://github.com/ACL-LLM-Research/BioLaySummarization

BioNLP 2025 Shared Tasks, pages 202–214
August 1, 2025 ©2025 Association for Computational Linguistics

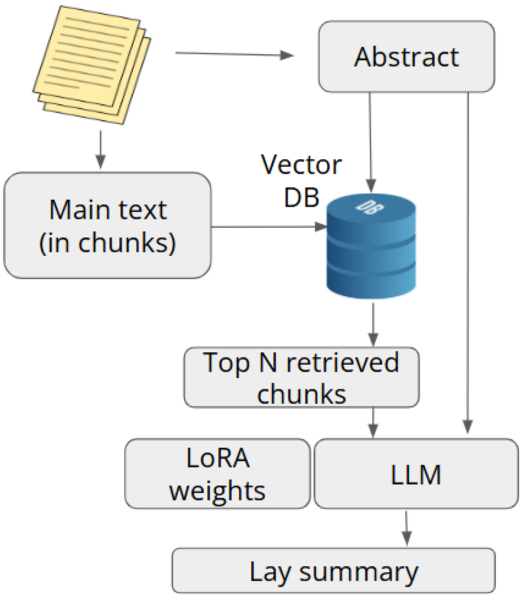

Figure 1: Overview of the proposed workflow for biomedical lay summarization. The abstract is used to query a vector database constructed from the segmented main text of the article. The retrieved content is then combined with the abstract and processed by a fine-tuned language model to generate a lay summary.

2 Methods

2.1 Datasets

In this study, we used a publicly available PLOS and eLife dataset (Goldsack et al., 2022), which includes both full research articles and their corresponding lay summaries written by the original authors or editors. Summary statistics of the data set can be found in the Appendix A.

2.2 Supervised Fine-Tuning

The LLaMA 3.1 8B model was used as the base model for supervised fine-tuning (Grattafiori et al., 2024). Given the size of the training set, we adopted Low-Rank Adaptation (LoRA), a parameter-efficient fine-tuning approach. A brief hyperparameter search was conducted based on the autoregressive loss. Further details are provided in the Appendix B.

2.3 Retrieval-Augmented Generation (RAG)

The vector database was constructed using the main text of each article. The text was segmented into 500-character chunks with a 50-character overlap. Each chunk was embedded using the all-MiniLM-L6-v2 model from the Sentence Transformers library, which encodes sentences and short paragraphs into dense vectors optimized for semantic similarity and retrieval (Reimers and Gurevych, 2019). The resulting embeddings were indexed using FAISS (Douze et al., 2025).

During the retrieval phase, each article's abstract was used as a query to retrieve semantically similar and contextually relevant content from the corresponding document in the vector database. The top five most relevant chunks, ranked by embedding similarity, were incorporated into the prompt alongside the original abstract.

2.4 Evaluation Metrics

We evaluated summary quality using three metric categories: relevance, readability, and factuality.

Relevance was evaluated using ROUGE (Lin, 2004), BLEU(Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005) and, BERTScore (Zhang et al., 2020), which quantify lexical and semantic overlap between the generated and reference summaries.

Readability was evaluated using the Flesch-Kincaid Grade Level (FKGL) (Kincaid et al., 1975), Dale-Chall Readability Score (DCRS) (Dale and Chall, 1948), Coleman-Liau Index (CLI) (Coleman and Liau, 1975) and LENS (Maddela et al., 2023).

Factuality metrics include AlignScore (Zha et al., 2023) and SummaC (Laban et al., 2022), which estimate the consistency of generated summaries with the source content.

Additionally, we explored the use of G-Eval, an LLM-based evaluator that provides a holistic assessment by jointly considering relevance, readability, and factuality. However, it requires further development and was not included in this study. Further details are provided in Appendix C.

3 Result

We first explored several strategies to improve model performance, including retrieval-augmented generation (RAG), LoRA-based fine-tuning, and prompt engineering using the validation set. Based on these evaluations, we selected the best-performing approach and compared its performance on the test set against that of general-purpose large language models.

3.1 Retrieval-Augmented Fine-Tuning

Our study utilizes LLaMA3.1-8B-Instruct as the primary baseline model. To assess the impact of scaling model size by an order of magnitude, we also included the LLaMA3-70B-Instruct base model. However, the performance of the 70B model was only marginally better than that of the 8B model (Table 1).

We then incorporated retrieval-augmented generation (RAG) to evaluate whether retrieved text chunks from the full text could enhance summary quality. The underlying hypothesis is that retrieved content may contain contextual information relevant to key points mentioned in the abstract, thereby providing additional background for generating more comprehensive and informative summaries. Compared to the base model, the RAG approach achieved higher scores in relevance metrics such as ROUGE and METEOR, as well as in most readability metrics across both datasets, although it underperformed in factuality metrics.

The main rationale for using retrieved text in place of the full article is to minimize computational overhead. We compared the performance of the RAG approach with models using full-text input. Interestingly, the results were dataset-dependent. On the PLOS dataset, the RAG-based model outperformed the full-text model across all relevance and factuality metrics. In contrast, on the eLife dataset, the RAG-based summaries underperformed relative to the full-text model in both relevance and readability metrics.(Table 1).

Next, we evaluated whether supervised fine-tuning using LoRA could enhance summary quality. Compared to the base model, the LoRA fine-tuned model demonstrated improvements across all evaluation metrics on both the PLOS and eLife datasets, with the exception of the readability metric LENS. (Table 1).

Finally, we assessed a combined approach using both LoRA and RAG to determine whether the two strategies are complementary. On the PLOS dataset, this combined model outperformed the base model as well as models using LoRA or RAG alone in both relevance and factuality metrics, though not in all readability metrics. On the eLife dataset, the combined model outperformed others in most relevance metrics and achieved higher scores in one factuality metric, AlignScore (Table 1).

3.2 Prompt-Based Trade-off between factuality and readability

Given the low scores observed in readability metrics such as FKGL, CLI, and DCRS, we modified the prompt instructions to enhance readability. Prompts 1 through 4 (B.2 to B.5) were designed to incrementally increase emphasis on readability, while progressively reducing focus on factual accuracy

The results from Prompt 1 to Prompt 4 without LoRA exhibit a consistent upward trend across all readability metrics (Table 2). However, this improvement in readability is accompanied by a decline in factuality, as evidenced by decreasing scores in AlignScore and SummaC. In contrast, for models fine-tuned using LoRA, the increase in readability is less consistent compared to models relying solely on RAG.

We also estimated the average readability metrics of the reference summaries using 100 examples from the eLife training set. These editor-written summaries achieved average scores of FKGL = 11.1694, CLI = 12.2691, DCRS = 11.1068, and LENS = 58.9321. In comparison, our generated summaries using the RAG approach with Prompt 4 produced slightly lower, but comparable results: FKGL = 11.6004, CLI = 12.5959, DCRS = 14.3942, and LENS = 52.1310.

3.3 Comparison against other pretrained LLMs

Given the strong overall performance of LLaMA 3.1–8B with retrieval-augmented fine-tuning using prompts that emphasize factual accuracy, we submitted its results for test set evaluation and compared them to summaries generated by various types of LLMs, including GPT, Qwen (a hybrid reasoning model), and Seed/Doubao (a mixture-of-experts model). The aggregated results across both datasets are presented in Table 3 .

Our model achieves the highest average scores in ROUGE, BLEU, METEOR, AlignScore, and SummaC, demonstrating superior performance compared to other LLMs, including GPT-4. Among the general-purpose systems, GPT-4 performs second-best on relevance metrics, but still lags behind our approach by over 0.02 in ROUGE and 0.04 in METEOR. In contrast, the Doubao and Qwen-3-32B models perform significantly worse, highlighting the effectiveness of retrieval-augmented LoRA fine-tuning for domain-specific summarization. In readability metrics, our system achieves stronger performance than GPT-3.5 on CLI, FKGL, and DCRS, although it underperforms relative to GPT-4.

Approach	Dataset	ROUGE	BLEU	METEOR	BERTScore	FKGL ↓	CLI ↓	DCRS ↓	LENS	AlignScore	SummaC
LLaMA3.1-8B base	PLOS	0.3015	6.5620	0.2523	0.8472	15.9551	14.3183	17.6025	43.7496	0.7888	0.6043
LLaMA3.1-70B base	PLOS	0.3177	5.9399	0.2732	0.8482	16.6955	14.6441	18.4586	56.0141	0.7838	0.6110
LLaMA3.1-8B +RAG	PLOS	0.3111	6.4173	0.2757	0.8448	15.6542	14.0322	16.3438	37.9543	0.7716	0.5969
LLaMA3.1-8B, full text	PLOS	0.2868	5.5198	0.2696	0.8386	**14.0785**	13.7034	**15.3931**	39.2586	0.7406	0.4899
LLaMA3.1-8B +LoRA	PLOS	0.3125	8.0553	0.2684	0.8483	14.2926	13.6650	15.7595	43.5402	0.7961	0.6606
LLaMA3.1-8B +RAG+LoRA	PLOS	**0.3682**	**13.1528**	**0.3294**	**0.8589**	16.0238	**13.6458**	16.2309	**59.2123**	**0.8905**	**0.8325**
LLaMA3.1-8B base	eLife	0.1938	0.9549	0.1247	0.8250	15.0991	14.1559	17.8718	**50.6078**	0.8171	0.5587
LLaMA3.1-70B base	eLife	0.2583	2.6717	0.2026	0.8237	16.2511	14.2092	17.7906	41.9114	0.8145	0.5141
LLaMA3.1-8B +RAG	eLife	0.2357	1.6102	0.1654	0.8208	14.7901	13.8905	16.3313	35.0491	0.7680	0.4821
LLaMA3.1-8B, full text	eLife	0.2475	2.7377	0.2267	0.8124	**12.8411**	13.7393	**14.9535**	16.7232	0.7919	0.4650
LLaMA3.1-8B +LoRA	eLife	0.2276	1.2622	0.1467	**0.8283**	14.2445	13.5854	16.2066	46.0314	0.8103	**0.5900**
LLaMA3.1-8B +RAG+LoRA	eLife	**0.3093**	**4.8882**	**0.2404**	0.8277	16.0863	**13.5323**	17.1463	49.9853	**0.8187**	0.5412

Table 1: Performance of models with RAG and LoRA on the validation set. ↓ Indicates that lower values correspond to better performance. Bold indicates the best score in each dataset. All metrics were computed on the full validation set (PLOS, $n = 1376$. eLife, $n = 271$).

Approach	Dataset	ROUGE	BLEU	METEOR	BERTScore	FKGL ↓	CLI ↓	DCRS ↓	LENS	AlignScore	SummaC
RAG, prompt 1	plos	0.3111	6.4173	0.2757	0.8448	15.6542	14.0322	16.3438	37.9543	0.7716	0.5969
RAG, prompt 2	plos	0.3139	6.3055	0.2856	0.8456	14.3461	13.6433	15.6403	46.0194	0.7479	0.5559
RAG, prompt 3	plos	0.3088	6.3005	0.2632	0.8492	13.0738	12.9733	14.2286	62.0774	0.7013	0.5466
RAG, prompt 4	plos	0.2966	4.5010	0.2493	0.8467	**11.7158**	**12.1700**	**12.5419**	66.4741	0.5951	0.5133
RAG+LoRA, prompt 1	plos	**0.3682**	**13.1528**	0.3294	**0.8589**	16.0238	13.6458	16.2309	59.2123	**0.8905**	**0.8325**
RAG+LoRA, prompt 2	plos	0.3485	9.8177	0.3227	0.8550	16.5375	13.5550	16.3920	66.1999	0.7753	0.6310
RAG+LoRA, prompt 3	plos	0.3601	10.1433	**0.3315**	0.8561	15.1839	13.3211	15.7658	69.0966	0.7598	0.5718
RAG+LoRA, prompt 4	plos	0.3434	8.2041	0.3230	0.8560	15.1830	12.8227	14.9630	**73.0403**	0.6322	0.5242
RAG, prompt 1	elife	0.2357	1.6102	0.1654	0.8208	14.7901	13.8905	16.3313	35.0491	0.7680	0.4821
RAG, prompt 2	elife	0.2638	2.2989	0.1846	0.8271	13.5047	13.3958	15.5217	46.3963	0.7802	0.5234
RAG, prompt 3	elife	0.2739	3.1185	0.2050	0.8283	11.8284	12.7894	14.3307	47.2025	0.7419	0.5303
RAG, prompt 4	elife	0.2771	3.2216	0.2031	0.8296	**11.6004**	**12.5959**	14.3942	52.1310	0.7366	0.5375
RAG+LoRA, prompt 1	elife	**0.3093**	4.8882	**0.2404**	0.8277	16.0863	13.5323	17.1463	49.9853	**0.8187**	**0.5412**
RAG+LoRA, prompt 2	elife	0.2886	4.4900	0.2186	0.8241	15.8230	15.0335	14.0296	52.6318	0.7171	0.5104
RAG+LoRA, prompt 3	elife	0.2957	4.8307	0.2215	0.8252	15.6024	14.8157	13.7995	52.2966	0.7302	0.5364
RAG+LoRA, prompt 4	elife	0.3061	**5.0620**	0.2317	**0.8303**	15.3697	13.9646	**13.3087**	**60.8793**	0.6404	0.4672

Table 2: The impact of prompt design on generated summaries using augmented LLaMA 3.1 models. Prompts 1 through 4 progressively increase emphasis on readability while reducing emphasis on factuality. ↓ indicates that lower values correspond to better performance. Bold values indicate the best score within each dataset. All metrics were computed on the full validation set (PLOS, $n = 1376$; eLife, $n = 271$).

Model	ROUGE	BLEU	METEOR	BERTScore	FKGL ↓	CLI ↓	DCRS ↓	LENS	AlignScore	SummaC
LLaMA3.1-8B +RAG+LoRA, prompt 1	**0.3469**	**8.6382**	**0.2978**	0.8534	16.9472	10.9176	17.2120	57.6922	**0.8801**	**0.7471**
LLaMA3.1-8B +RAG, prompt 4	0.2985	4.6963	0.2499	0.8457	12.9965	10.3171	14.5694	53.3393	0.7646	0.5704
Seed/Doubao-1.5-pro, RAG, prompt 1	0.1371	0.4055	0.1202	0.8052	12.3599	11.0582	15.6888	71.4021	0.3423	0.4382
Qwen3-32B, RAG, prompt 1	0.1926	1.4937	0.1396	0.8338	16.4236	14.2241	19.7064	40.6607	0.6860	0.5315
GPT3.5, RAG, prompt 1	0.2918	3.9624	0.2076	0.8536	17.5771	12.1847	18.9784	66.3074	0.8047	0.5118
GPT3.5, RAG, prompt 4	0.2543	2.2707	0.1709	0.8544	14.7574	11.7962	16.9538	74.9194	0.7850	0.5180
GPT4, RAG, prompt 4	0.3207	5.4428	0.2532	**0.8554**	**12.2789**	**9.5065**	**13.3833**	**80.4591**	0.6754	0.5210

Table 3: Final submission and test set performance compared to other general-purpose LLMs. The table reports average results across the PLOS and eLife datasets. ↓ indicates that lower values correspond to better performance. Bold values indicate the best scores. All metrics were computed on the test set (PLOS, $n = 142$. eLife, $n = 142$).

4 Discussion and Conclusion

Applying both LoRA and RAG to fine-tune LLaMA3.1 resulted in superior overall performance on the biomedical lay summarization task compared to using the base model or applying LoRA or RAG individually. This combined approach substantially improved relevance and factuality metrics, though it slightly reduced performance on most readability metrics. The gains in relevance and factuality are likely attributable to the additional contextual information retrieved from the full text, which often contains factual content present in the reference summaries but absent from the abstract. The slight decline in readability metrics, such as FKGL, CLI, and DCRS, may result from the introduction of new concepts via the retrieved content or from the integration of additional information using more complex sentence structures, such as subordinate clauses.

Our prompt design experiments revealed a trade-off between factuality and readability in LLM-generated summaries, suggesting that efforts to simplify language or meet brevity constraints may compromise the accurate representation of complex scientific content. It may be challenging for a single model to simultaneously enforce simplified vocabulary and sentence structures, comply with word count constraints, and extract essential information while preserving technical precision. A potential solution is to adopt a two-stage summarization framework (Goldsack et al., 2025), where an "author" model first extracts key factual content, followed by a "writer" model that generates a more readable summary while preserving that information.

Our RAG fine-tuned LLaMA3.1 model demonstrated superior performance in relevance and factuality metrics compared to pretrained general-purpose LLMs in this summarization task. However, we also observed that the pretrained GPT-4 model excels in readability metrics while maintaining competitive performance in relevance and factuality. This suggests that GPT-4 may serve as a strong base model for fine-tuning, potentially achieving well-balanced performance across all evaluation criteria, as demonstrated in previous work (You et al., 2024). Nevertheless, leveraging GPT-4 for fine-tuning and inference entails significantly higher computational and financial costs.

Limitations

This study has several limitations. First, the RAG component relied exclusively on the main text of each article. As a result, it may have omitted essential background information, such as fundamental biological concepts, which are critical for generating accurate and accessible lay summaries. Future work could enhance summary quality by incorporating external domain-specific resources—such as biomedical ontologies or reference texts—into the RAG pipeline. Additionally, the embedding model used in our RAG implementation was a small, cost-efficient variant. Employing larger and more powerful models, such as all-mpnet-base-v2, may further improve retrieval quality and overall summarization performance. Third, we used prompt templates optimized for LLaMA 3.1 to evaluate other LLMs, which may disadvantage models whose optimal prompts differ in structure or emphasis.

Acknowledgments

We are grateful for Peter Grabowski's insightful feedback regarding the LLM based evaluation and experimental design.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Siyu Bao, Ruijing Zhao, Siqin Zhang, Jinghui Zhang, Weiyin Wang, and Yunian Ru. 2024. Ctyun AI at BioLaySumm: Enhancing lay summaries of biomedical articles through large language models and data augmentation. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 837–844, Bangkok, Thailand. Association for Computational Linguistics.

Meri Coleman and T. L. Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283–284.

Edgar Dale and Jeanne S. Chall. 1948. A formula for predicting readability: Instructions. *Educational Research Bulletin*, 27(1):37–54.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2025. The faiss library. *Preprint*, arXiv:2401.08281.

Tomas Goldsack, Zheheng Luo, Qianqian Xie, Carolina Scarton, Matthew Shardlow, Sophia Ananiadou, and Chenghua Lin. 2023. 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, and Chenghua Lin. 2025. Leveraging large language models for zero-shot lay summarisation in biomedicine and beyond. *Preprint*, arXiv:2501.05224.

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.

Rita González-Márquez, Luca Schmidt, Benjamin M. Schmidt, Philipp Berens, and Dmitry Kobak. 2024. The landscape of biomedical research. *Patterns*, 5(6):100968.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

J. Peter Kincaid, Robert P. Fishburne, 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. Technical Report Research Branch Report 8-75, Naval Technical Training Command, Millington TN Research Branch.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

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.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Chenghao Xiao, Kun Zhao, Xiao Wang, Siwei Wu, Sixing Yan, Tomas Goldsack, Sophia Ananiadou, Noura Al Moubayed, Liang Zhan, William Cheung, and Chenghua Lin. 2025. Overview of the biolaysumm 2025 shared task on lay summarization of biomedical research articles and radiology reports. In *The 24th Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Vienna, Austria. Association for Computational Linguistics.

Zhiwen You, Shruthan Radhakrishna, Shufan Ming, and Halil Kilicoglu. 2024. UIUC_BioNLP at BioLaySumm: An extract-then-summarize approach augmented with Wikipedia knowledge for biomedical lay summarization. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 132–143, Bangkok, Thailand. Association for Computational Linguistics.

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.

Haopeng Zhang, Philip S. Yu, and Jiawei Zhang. 2024. A systematic survey of text summarization: From statistical methods to large language models. *Preprint*, arXiv:2406.11289.

Dataset	Train	Validation	Test
PLOS	24,773	1,376	142
eLife	4,346	242	142

Table 4: Number of examples in the training, validation, and test sets of the PLOS and eLife lay summary dataset.

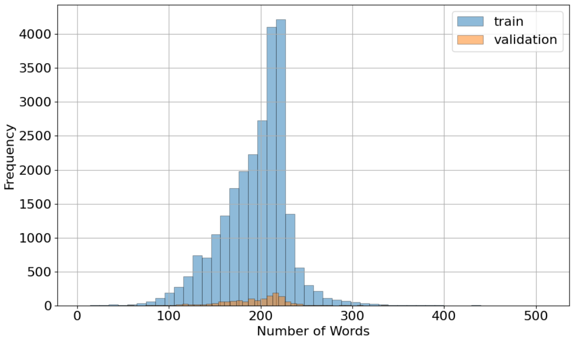

Figure 2: Word counts of PLOS reference summaries in the training and validation sets.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

A Dataset Summary Statistics

The dataset was divided into training, validation, and test sets, as shown in Table 4.

Summary statistics of word counts are presented in Tables 5 and 6 to confirm that the validation and test splits are representative of the dataset. The training, validation, and test sets display comparable mean and median word counts. However, 13 instances in the PLOS training set contain incorrectly phrased abstracts, each comprising fewer than 500 tokens according to the LLaMA 3 tokenizer. These instances were identified as having improperly parsed abstracts and were removed prior to training.

The reference summaries typically range from 100–300 words for PLOS and 200–600 words for eLife (Figure 2 and Figure 3). These ranges informed the prompt design, enabling the model to generate summaries of comparable lengths .

B Fine-Tuning

The prompts used for LoRA fine-tuning—with and without RAG—are provided in B.1 and B.2, respectively. Similar prompts were used during inference, except that the reference summary part was omitted.

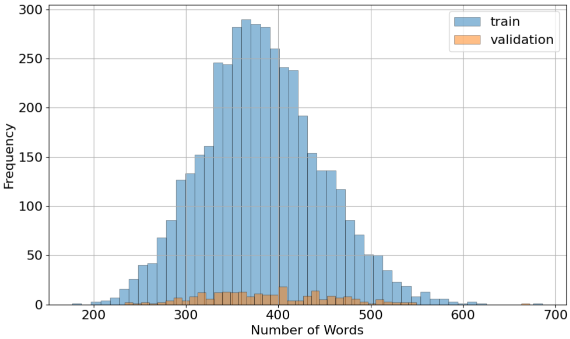

Figure 3: Word counts of eLife reference summaries in the training and validation sets.

The final LoRA configuration employed a LoRA rank of 8, a LoRA alpha of 16, a LoRA dropout of 0.1, and a learning rate of 1×10^{-5}. Hyperparameters such as the LoRA rank (8, 16) were explored using the PLOS dataset, but performance differences were minimal. The number of epochs (1, 2, 4) was also explored. The optimal number of training epochs was found to be 1 for the PLOS dataset and 2 for the eLife dataset. The training and validation loss curves are available in the GitHub repository.

A series of prompts with progressively increased emphasis on readability were explored: Prompt 1 (B.2), Prompt 2 (B.3), Prompt 3 (B.4), and Prompt 4 (B.5). Blue color coding indicates instructions related to accuracy, while orange color coding highlights instructions aimed at improving readability.

C Evaluation of G-Eval

Multiple classical metrics were employed in this study, some of which exhibited contradictory behavior during prompt optimization. This highlights the challenge of determining appropriate weights for each metric in order to construct a meaningful overall evaluation score. Previously, equal weights were assigned to each metric to calculate average performance within each evaluation aspects (Goldsack et al., 2024). Recent advancements have introduced the "LLM-as-a-Judge" paradigm, wherein large language models are employed as evaluators for complex tasks, offering scalable, cost-effective, and consistent assessments across diverse domains (Gu et al., 2025).

In this preliminary study, we employed G-Eval, an LLM-based evaluation framework that prompts a language model to assign scores and provide justifications based on criteria such as relevance, readability, and factuality (Liu et al., 2023). G-Eval

	Text	Min	Max	Mean	Median
Train	Abstract	71	509	166	165
	Main text	324	28,696	10,200	9.890
Validation	Abstract	76	306	166	165
	Main text	3,408	23,048	10,031	9,707
Test	Abstract	83	464	267	220
	Main text	2,666	16,954	8,157	8,032

Table 5: Words counts of abstracts and articles from the eLife dataset.

	Text	Min	Max	Mean	Median
Train	Abstract	2*	701	268	269
	Main text	748	26,643	6,754	6.581
Validation	Abstract	93	561	271	273
	Main text	933	24,751	8,869	8,649
Test	Abstract	97	377	245	245
	Main text	3,316	17,330	7,735	7,521

Table 6: Word counts of abstracts and articles from the PLOS dataset. * indicate instances with unusually low word counts due to incorrectly parsed abstracts, which were removed prior to training.

was implemented using the GPT-3.5-turbo model, and the evaluation criteria are detailed in Box C.1.

To evaluate the effectiveness of G-Eval for this task, we conducted a controlled experiment using synthesized data. Specifically, we examined whether G-Eval scores could differentiate among positive controls, negative controls, and standard summaries generated with the LLaMA3.1 model using various prompts. The positive controls (Paraphrased Gold Summaries) were created by paraphrasing the reference summaries to preserve their factual content while altering surface form. The negative controls (Intentionally Degraded Summaries) were generated by prompting the model to produce outputs characterized by vague language, poor structure, and incorrect terminology. The standard summaries were generated directly from abstracts using a conventional prompt. All prompts used to generate this synthetic data are listed in Table 7.

The results showed that paraphrased reference summaries achieved the highest median G-Eval scores, while intentionally degraded summaries received the lowest scores. Summaries generated using the standard prompt fell between these two extremes (Figure 4). These findings suggest that G-Eval is effective in distinguishing between summaries of varying quality.

In addition, summaries that received low G-Eval scores were manually reviewed to assess the justifications provided by the G-Eval framework for their evaluation.

We applied G-Eval scoring to 20 test set examples across four model configurations: the LLaMA 3.1 baseline, LLaMA 3.1 + LoRA, LLaMA 3.1 + RAG, and LLaMA 3.1 + RAG + LoRA (Figure 5). Consistent with the results in Table 1, the model fine-tuned with both LoRA and RAG achieved the highest median G-Eval score, suggesting that G-Eval is capable of distinguishing higher-performing models from lower-performing ones. However, the boxplot reveals substantial variance across the 20 evaluated samples, indicating that a larger sample size would be necessary to establish statistical significance.

Upon reviewing examples with low G-Eval scores, we found that the most common reason for low performance was the omission of key details present in the reference summary. This issue likely stems from limitations in the abstract, which may lack sufficient context, and from retrieved chunks that failed to supplement the missing information.

Box B.1: LoRA without RAG, prompt 1

```
#system:You are an expert science communicator. Your task is to generate a clear,
accurate, and formal summary of biomedical research articles.
The summary should be accessible to a general audience while maintaining
scientific rigor.
#user:
Title: (...)
Abstract: (...)

Provide a formal summary of the article in {summary_word_len} words.
Do not include explanations, self-reflections, or additional notes.
Keep the response strictly to the summary. The output should begin directly with
the summary text itself.

#assistant:
(ref summary...)
```

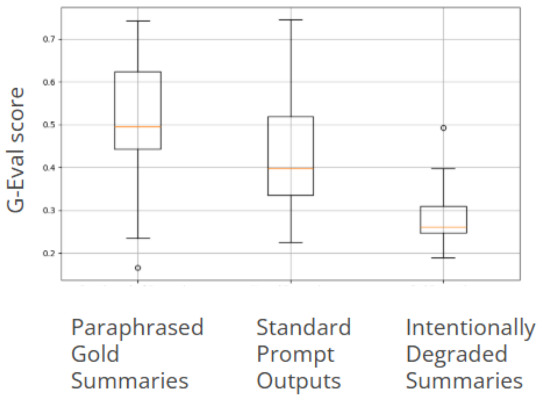

Figure 4: G-Eval scores for summaries generated by the LLaMA3.1-Instruct model and control conditions. Paraphrased Gold Summaries were created by rephrasing the original lay summaries while preserving their meaning. Intentionally Degraded Summaries were generated by explicitly prompting LLaMA 3.1 to produce outputs with vague language, poor structure, and incorrect terminology. n=20

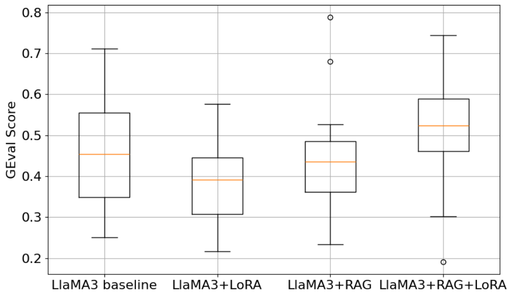

Figure 5: G-Eval scores for summaries generated by the LLaMA 3.1 model and fine-tuned models, with or without RAG. n=20

Box B.2: LoRA with RAG, prompt 1

```
#system: You are an expert science communicator. Your task is to generate a
clear, accurate, and formal summary of biomedical research articles.
The summary should be accessible to a general audience while maintaining
scientific rigor.
#user:
Title: (...)
Abstract: (...)

Supporting Text:
(retrieved text chunks...)

Provide a formal summary of the article in {summary_word_len} words.
Do not include explanations, self-reflections, or additional notes.
Keep the response strictly to the summary. The output should begin directly with
the summary text itself.

#assistant:
(ref summary...)
```

Box B.3: LoRA with RAG, prompt 2

```
#system: You are an expert science communicator. Your task is to generate a
clear, accurate, and formal summary of biomedical research articles.
The summary should be accessible to a general audience using plain language,
short sentences, and avoiding technical jargon where possible, while maintaining
scientific accuracy.
#user:
Title: (...)
Abstract: (...)

Supporting Text:
(retrieved text chunks...)

Provide a formal summary of the article in {summary_word_len} words.
Do not include explanations, self-reflections, or additional notes.
Keep the response strictly to the summary. The output should begin directly with
the summary text itself.

#assistant:
(ref summary...)
```

Box B.4: LoRA with RAG, prompt 3

```
#system: You are an expert science communicator. Your task is to generate a
clear, accurate, and formal summary of biomedical research articles.
The summary should be accessible to a general audience. Use simple sentence
structures, common words, and avoid long or complex clauses. Aim for a tone
similar to science communication articles in outlets like Scientific American or
NIH press releases.
#user:
Title: (...)
Abstract: (...)

Supporting Text:
(retrieved text chunks...)

Provide a formal summary of the article in {summary_word_len} words.
Do not include explanations, self-reflections, or additional notes.
Keep the response strictly to the summary. The output should begin directly with
the summary text itself.

#assistant:
(ref summary...)
```

Box B.5: LoRA with RAG, prompt 4

```
#system: You are an expert science communicator. Your task is to generate a
summary of biomedical research articles.
The summary should be accessible to a general audience. Use simple sentence
structures, common words, and avoid long or complex clauses. Aim for a tone
similar to science communication articles in outlets like Scientific American or
NIH press releases.
#user:
Title: (...)
Abstract: (...)

Supporting Text:
(retrieved text chunks...)

Provide a formal summary of the article in {summary_word_len} words.
Do not include explanations, self-reflections, or additional notes.
Keep the response strictly to the summary. The output should begin directly with
the summary text itself.

#assistant:
(ref summary...)
```

Box C.1: G-Eval Evaluation Criteria

Evaluate the generated lay summary on the following three criteria: 1. **Relevance (1–5)**: Does the summary retain all major findings and themes of the source abstract? Score higher if it covers key points, even if phrased differently. Penalize only if essential information is missing or incorrect topics are introduced. 2. **Readability (1–5)**: Is the summary easy to understand for a non-expert audience? Consider fluency, sentence structure, and clarity. Avoid penalizing for simplified language unless it introduces confusion. 3. **Factuality (1–5)**: Does the summary accurately reflect the scientific claims in the source abstract? Check for hallucinations or misinterpretations, not just omissions. Each criterion should be scored from 1 (poor) to 5 (excellent). Then provide a final **Overall Score**.

Synthetic Data Type	Prompt

Paraphrased Gold Summaries	`#system`
	You are a professional science communicator. Your role is to paraphrase lay summaries with precision, maintaining the original meaning and content without introducing interpretation or additional information.
	`#user`
	Title: (...)
	Summary: (...)
	Rephrase the summary in 100–300 words. Do not include explanations, commentary, or additional remarks.
	Keep the response strictly to the summary.
	`#assistant`

Intentionally Degraded Summaries	`#system`
	You are a deliberately ineffective science communicator. Your task is to generate an example of a poorly written summary of biomedical research. This summary should reflect common mistakes in science communication, such as vague language, poor structure, and misuse of terminology. The summary may also include minor factual inaccuracies or exaggerated claims to illustrate how misleading summaries might appear. This output will be used strictly for educational comparison with well-written summaries.
	`#user`
	Title: (...)
	Abstract: (...)
	Provide a poor-quality summary of the article in 100–300 words, reflecting issues like lack of clarity, overgeneralization, or scientific inaccuracy (intended for contrastive purposes only). At least some summary needs to be generated. Do not include explanations, self-reflections, or additional notes.
	Keep the response strictly to the summary.
	`#assistant`

Standard Prompt Outputs	`#system`
	You are an expert science communicator. Your task is to generate a clear, accurate, and formal summary of biomedical research articles. The summary should be accessible to a general audience while maintaining scientific rigor.
	`#user`
	Title: (...)
	Abstract: (...)
	Provide a formal summary of the article in 100–300 words.
	Do not include explanations, self-reflections, or additional notes.
	Keep the response strictly to the summary.
	`#assistant`

Table 7: Prompts used to generate paraphrased, degraded, and standard summaries for evaluating G-Eval.