MedSummRAG: Domain-Specific Retrieval for Medical Summarization

Guanting Luo
The University of Osaka
guanting.luo@ist.osaka-u.ac.jp

Yuki Arase
Institute of Science Tokyo
arase@c.titech.ac.jp

Abstract

Medical text summarization faces significant challenges due to the complexity and domain-specific nature of the language. Although large language models have achieved significant success in general domains, their effectiveness in the medical domain remains limited. This limitation stems from their insufficient understanding of domain-specific terminology and difficulty in interpreting complex medical relationships, which often results in suboptimal summarization quality. To address these challenges, we propose MedSummRAG, a novel retrieval-augmented generation (RAG) framework that integrates external knowledge to enhance summarization. Our approach employs a fine-tuned dense retriever, trained with contrastive learning, to retrieve relevant documents for medical summarization. The retrieved documents are then integrated with the input text to generate high-quality summaries. Experimental results show that MedSummRAG achieves significant improvements in ROUGE scores on both zero/few-shot and fine-tuned language models, outperforming baseline methods. These findings underscore the importance of RAG and domain adaptation of the retriever for medical text summarization. The source code of this paper can be obtained from: https://github.com/guantingluo98/MedSummRAG

1 Introduction

Medical text summarization is a crucial task for helping medical practitioners and patients, aiming to distill complex and information-dense medical documents into concise, accurate, and clinically useful summaries (Xie et al., 2023). This task is particularly challenging due to the specialized nature of medical language and the presence of domain-specific terminologies (Chaves et al., 2022). Traditional summarization models often struggle in this domain, as they may fail to generate satisfactory summaries.

With the rise of large language models (LLMs), significant advancements have been made in general-domain summarization (Pu et al., 2023). However, medical summarization presents unique challenges, such as domain-specific terminology and complex relationships, which generic LLMs struggle to address effectively. LLMs trained on broad-domain corpora tend to overlook key medical concepts, misinterpret medical abbreviations, and produce hallucinated content that could mislead practitioners and researchers (Li et al., 2024; Hosseini et al., 2024). These limitations highlight the need for models that can effectively incorporate external domain knowledge. By leveraging external knowledge documents, such as healthcare question-answer pairs, models can better understand domain-specific concepts, reduce errors, and generate high-quality summaries.

In this work, we propose MedSummRAG (Medical Summarization with Retrieval-Augmented Generation), a novel retrieval-augmented generation (RAG) framework designed specifically for medical text summarization. By leveraging external medical knowledge, MedSummRAG enhances the quality of generated summaries. Our approach employs a fine-tuned dense retriever, trained using contrastive learning (van den Oord et al., 2019), to effectively identify domain-relevant documents.

The key contribution of our work is the novel RAG framework for medical text summarization. Our approach improves retrieval quality by leveraging contrastive learning that employs synthetic positive samples generated using an LLM. This enables the framework to effectively identify domain-relevant documents, improving the overall quality of generated summaries. We conduct experiments to investigate the effectiveness of MedSummRAG. Our results demonstrate consistent improvements measured by ROUGE scores in multiple configurations: both on zero/few-shot and fine-tuned language models.

Proceedings of the 24th Workshop on Biomedical Language Processing (BioNLP 2025), pages 27–33

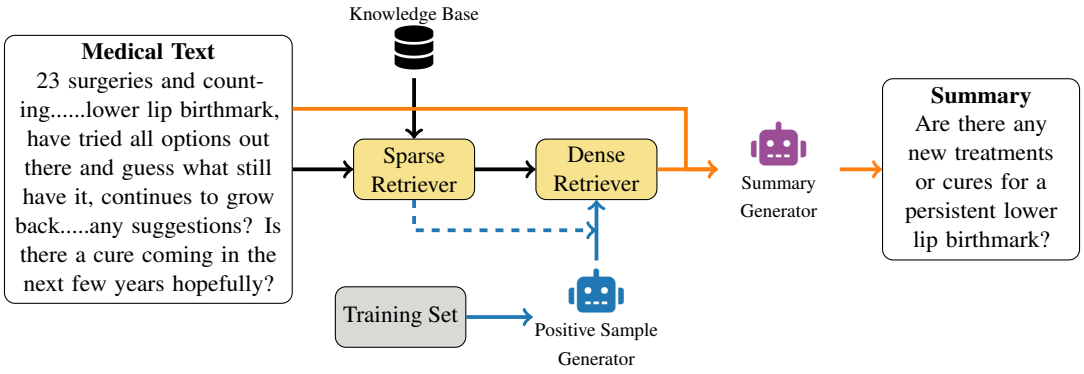

Figure 1: Overview of our MedSummRAG framework. Black arrows indicate retrieving relevant documents by sparse and dense retrievers. Blue dashed arrow represents the negative sample extraction, while solid blue arrows show the generation of synthetic positive samples. Both positive and negative samples are used to fine-tune the dense retriever. Orange arrows show the concatenation of medical text and the retrieved document as input to the summary generator, producing the summary.

2 Related Work

Medical summarization has been a long-standing research problem due to its critical role in supporting clinical decision-making and healthcare planning. With the rise of pre-trained language models, significant progress has been made in medical summarization. Pre-trained language models have demonstrated the ability to generate medical summaries, such as doctor-patient conversation summaries, by utilizing knowledge derived from pretrained models (Zhang et al., 2021). Balde et al. (2024) proposed MEDVOC, a dynamic vocabulary adaptation strategy that optimizes pre-trained language models' vocabulary for medical text summarization, achieving improvements in high Out-Of-Vocabulary settings.

Despite the progress enabled by pre-trained models in medical summarization, their pre-trained knowledge may be insufficient for handling specific downstream tasks. RAG offers a promising solution by integrating external knowledge to enhance the overall quality of generations (Fan et al., 2024). Recent studies have demonstrated the potential of RAG in various domains, such as decision-making tasks (Lee et al., 2024); question answering (Jeong et al., 2024); and radiology report generation (Xia et al., 2024). Although RAG has demonstrated success in various domains, its application to medical summarization remains underexplored. Our work aims to propose a RAG framework specifically adapted for medical summarization to generate high-quality summaries.

3 Proposed Method

The overall workflow of our approach is illustrated in Figure 1. The proposed method consists of document retrieval (Section 3.1) and summary generation (Section 3.2). For improving the retrieval step to adapt to the medical domain, we employ contrastive learning with synthetic data (Section 3.3).

3.1 Document Retrieval

We employ the BM25 (Robertson et al., 2009; Lù, 2024) ranking function to retrieve an initial set of candidate documents based on lexical similarity to the input text. This sparse retrieval method serves to reduce the computational cost of subsequent dense retrieval by narrowing down the search space to a manageable set of candidate documents.

A dense retriever then re-ranks the highly-ranked documents retrieved by the sparse retriever and selects the most relevant document for the generation stage. This step should ensure that the retrieved document is lexically and semantically aligned with the input text to provide useful knowledge for summarization.

3.2 Generation

The generation stage produces summaries based on the input text and the retrieved document. Following the approach of Lewis et al. (2020), we simply concatenate the retrieved document with the input text and feed the combined input into a language model. The generator is expected to produce coherent and factually accurate summaries, leveraging both the input text and the external knowledge provided by the retrieved document.

You're a retrieval augmented generation assistant, skilled in generating retrieval targets for auto summarization via RAG. Here is the input-summary pair from a training set: INPUT: {Train set input text} SUMMARY: {Train set summary} Please help me with generating one fake retrieved question-answer document that would be useful for training a dense retrieval model for automatic summarization via retrieval augmented generation. The fake retrieved document should have this kind of format: QUESTION: ANSWER:

Table 1: Promt for synthetic sample generation

3.3 Domain Adaptation of Retriever

The retrieval stage aims to identify the most relevant document from a knowledge base to enhance the summarization process. Although existing RAG methods have shown great success in question-answering tasks (Asai et al., 2023; Xiong et al., 2024), they often struggle to identify documents that are truly useful for medical text summarization. This is because pre-trained dense retrievers lack the ability to understand what document structures and content are beneficial for enhancing summarization in the medical domain.

The key challenge in fine-tuning retrievers for medical summarization is the lack of training data. To address this challenge, we leverage an LLM to generate synthetic positive samples that capture the structural and contextual patterns useful for summarization. Specifically, for each text-summary pair in a training set of summarization, we prompt the LLM to generate a synthetic pair that may enhance medical summarization. Table 1 shows the prompt we used.

We then fine-tune the dense retriever using contrastive learning, inspired by the work of Huang et al. (2023), which improves its ability to retrieve documents relevant to medical summarization. For negative samples, we randomly select low-ranked documents by the sparse retriever that should be less relevant to the input text. We optimize the retrieval model using the InfoNCE loss (van den Oord et al., 2019).

4 Experiments

4.1 Evaluation Dataset & Knowledge Base

Evaluation Dataset: We use the CHQ-Summ dataset (Yadav et al., 2022), which consists of consumer health questions formulated by non-experts, paired with brief summaries of the corresponding questions. The questions are sourced from Yahoo! Answers L6 corpus[1]. The dataset contains $1,000$ training samples, 107 validation samples, and 400 test samples. We evaluate the performance of our method using ROUGE (Lin, 2004) scores, including ROUGE-1, ROUGE-2, and ROUGE-L.

Knowledge Base: We construct the knowledge base using Yahoo! Answers L6 corpus, which contains more than 4 million question-answer pairs. Each document in the knowledge base represents a single question-answer pair. The content covered in this corpus extends far beyond the scope of healthcare and medicine, encompassing a wide range of topics. To prevent data leakage, we exclude all question-answer pairs that overlap with the CHQ-Summ dataset.

4.2 Implementation Details

We employed BM25 for sparse retrieval, which retrieved the top 150 documents for each input text. We employed the BGE-M3 (Chen et al., 2024) model as the base dense retriever. For contrastive learning, the positive samples were generated by a frozen Qwen-2.5-7B-Instruct model[2], while the negative samples were constructed by randomly sampling 3 documents from the BM25-ranked documents in the range of positions 101 to 150 for each training sample. The BGE-M3 model was fine-tuned for 5 epochs with a total batch size of 8. After fine-tuning, the BGE-M3 model re-ranked the top 20 documents retrieved by BM25 and selected the top 1 document for the generator.

4.3 Experiment Settings

To evaluate the effectiveness of our MedSumm-RAG approach, we conducted four sets of experiments with different generator settings: standard fine-tuning, few-shot prompting, and Low Rank Adapters (LoRA) (Hu et al., 2022) fine-tuning on different language models. Specifically, we employed (1) BioBART-large (Yuan et al., 2022): the

[1] https://webscope.sandbox.yahoo.com/catalog.php?datatype=l&did=11

[2] https://qwenlm.github.io/blog/qwen2.5/

Base Model	Setting	ROUGE-1	ROUGE-2	ROUGE-L
BioBART-large (Standard Fine-tuned)	Baseline	41.22	23.17	38.79
	+ Naive RAG	42.19	22.95	38.79
	+ Fine-tuned RAG	**44.50**	**24.58**	**41.19**
Qwen-2.5-7B-Instruct (1-shot Prompting)	Baseline	34.97	13.85	32.82
	+ Naive RAG	38.53	16.42	33.61
	+ Fine-tuned RAG	**39.45**	**17.59**	**34.60**
Qwen-2.5-7B-Instruct (2-shot Prompting)	Baseline	38.15	16.34	33.82
	+ Naive RAG	39.89	18.00	35.28
	+ Fine-tuned RAG	**40.27**	**18.30**	**35.95**
Qwen-2.5-7B-Instruct (LoRA Fine-tuned)	Baseline	42.21	21.99	38.84
	+ Naive RAG	42.56	21.80	39.32
	+ Fine-tuned RAG	**42.95**	**22.82**	**40.03**

Table 2: Performance comparison of different base models on the CHQ-Summ dataset. Results demonstrate the effectiveness of our method across various models, few-shot scenarios, and fine-tuning strategies.

model has shown its strong performance in medical text generation tasks. BioBART-large was first fine-tuned using the training set without RAG, followed by the second stage of fine-tuning with RAG. Each fine-tuning consisted of 20 epochs with a batch size of 8. We also experimented with (2) Qwen-2.5-7B-Instruct with One-Shot Prompting, (3) Qwen-2.5-7B-Instruct with Two-Shot Prompting, and (4) Qwen-2.5-7B-Instruct with LoRA Fine-Tuning (Hu et al., 2022): the model was fine-tuned using LoRA for 10 epochs with a batch size of 8. LoRA fine-tuning was performed with a rank of 8, alpha of 16, and no dropout. The details of the prompts are described in example A.1 and example A.2

In all settings, the baseline is the corresponding fine-tuned model or a few-shot prompted models without RAG. In addition, we also compared to a naive RAG where the retriever has not been fine-tuned, i.e., without domain adaptation. All the experiment was conducted on a single NVIDIA A6000 48G GPU.

4.4 Results

In this section, we highlight the key contribution of our RAG-enhanced approach, demonstrating its effectiveness across different models, few-shot settings, and fine-tuning strategies. A consistent performance gap between naive RAG and fine-tuned RAG underscores the importance of domain-adaptive retrieval. This contrast suggests that synthetic examples play a key role in improving the relevance of the retrieval and the overall quality of the summary.

For the standard fine-tuned BioBART-large model, our method significantly improves performance. With naive RAG, only the ROUGE-1 score shows a marginal improvement, while the ROUGE-2 score slightly decreases, and the ROUGE-L score remains unchanged. However, with MedSumm-RAG, the BioBART-large model achieves a notable increase in ROUGE scores, highlighting the effectiveness of integrating external knowledge through domain-adapted retriever.

For the Qwen-2.5-7B-Instruct model in few-shot prompt settings, our method consistently enhances performance without fine-tuning the generator. Even with naive RAG, we observe modest improvements in ROUGE scores. Fine-tuning the RAG component further boosts performance, demonstrating the effectiveness of our method even when the generator is frozen. Additional prompt examples also contribute to improved results.

For the LoRA fine-tuned Qwen-2.5-7B-Instruct model, integrating naive RAG yields marginal improvements in ROUGE-1 and ROUGE-L, while ROUGE-2 experiences a slight decline compared to the baseline. In contrast, our domain-adapted RAG enhances performance across all ROUGE metrics, demonstrating the importance of optimizing the retrieval process to effectively leverage external knowledge in the LoRA fine-tuning setting.

5 Conclusion

Our experimental results highlight the effectiveness of leveraging external knowledge for adapting language models to medical summarization tasks, addressing the challenge of domain adaptation in specialized medical contexts. Future work includes extending our approach to a larger-scale knowledge base to further enhance retrieval effectiveness. Additionally, beyond ROUGE-based evaluation, in-

corporating human evaluation could provide deeper insights into the quality of generated summaries. Furthermore, exploring the application of our fine-tuned RAG framework to other medical summarization tasks, such as radiology report summarization, is another promising direction for advancing our work.

Limitations

While our proposed method demonstrates promising results in improving medical text summarization, its generalizability remains to be validated. Our experiments are conducted exclusively on the CHQ-Summ dataset, which focuses on summarizing customer health questions. While this dataset provides a valuable benchmark for medical question summarization, it does not fully represent the diversity of medical texts, such as clinical notes, or discharge summaries. In addition, while the Yahoo! Answers L6 corpus offers broad coverage, it may contain content of varying accuracy, which motivates future exploration of more medically curated sources to further reduce hallucination risks.

Acknowledgements

This work was supported by Cross-ministerial Strategic Innovation Promotion Program (SIP) on "Integrated Health Care System" Grant Number JPJ012425.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *Preprint*, arXiv:2310.11511.

Gunjan Balde, Soumyadeep Roy, Mainack Mondal, and Niloy Ganguly. 2024. Medvoc: vocabulary adaptation for fine-tuning pre-trained language models on medical text summarization. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 6180–6188.

Andrea Chaves, Cyrille Kesiku, and Begonya Garcia-Zapirain. 2022. Automatic text summarization of biomedical text data: a systematic review. *Information*, 13(8):393.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6491–6501.

Manda Hosseini, Mandana Hosseini, and Reza Javidan. 2024. Leveraging large language models for clinical abbreviation disambiguation. *Journal of medical systems*, 48(1):27.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Wenyu Huang, Mirella Lapata, Pavlos Vougiouklis, Nikos Papasarantopoulos, and Jeff Pan. 2023. Retrieval augmented generation with rich answer encoding. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1012–1025, Nusa Dua, Bali. Association for Computational Linguistics.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. 2024. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7036–7050, Mexico City, Mexico. Association for Computational Linguistics.

Myeonghwa Lee, Seonho An, and Min-Soo Kim. 2024. PlanRAG: A plan-then-retrieval augmented generation for generative large language models as decision makers. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6537–6555, Mexico City, Mexico. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Songda Li, Yunqi Zhang, Chunyuan Deng, Yake Niu, and Hui Zhao. 2024. Better late than never: Model-agnostic hallucination post-processing framework towards clinical text summarization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 995–1011, Bangkok, Thailand. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summariza-*

tion Branches Out, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Xing Han Lù. 2024. Bm25s: Orders of magnitude faster lexical search via eager sparse scoring. *Preprint*, arXiv:2407.03618.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead. *Preprint*, arXiv:2309.09558.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation learning with contrastive predictive coding. *Preprint*, arXiv:1807.03748.

Peng Xia, Kangyu Zhu, Haoran Li, Hongtu Zhu, Yun Li, Gang Li, Linjun Zhang, and Huaxiu Yao. 2024. RULE: Reliable multimodal RAG for factuality in medical vision language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1081–1093, Miami, Florida, USA. Association for Computational Linguistics.

Qianqian Xie, Zheheng Luo, Benyou Wang, and Sophia Ananiadou. 2023. A survey for biomedical text summarization: From pre-trained to large language models. *Preprint*, arXiv:2304.08763.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251, Bangkok, Thailand. Association for Computational Linguistics.

Shweta Yadav, Deepak Gupta, and Dina Demner-Fushman. 2022. Chq-summ: A dataset for consumer healthcare question summarization. *Preprint*, arXiv:2206.06581.

Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. 2022. BioBART: Pretraining and evaluation of a biomedical generative language model. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 97–109, Dublin, Ireland. Association for Computational Linguistics.

Longxiang Zhang, Renato Negrinho, Arindam Ghosh, Vasudevan Jagannathan, Hamid Reza Hassanzadeh, Thomas Schaaf, and Matthew R. Gormley. 2021. Leveraging pretrained models for automatic summarization of doctor-patient conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3693–3712, Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Appendix

This appendix shows the prompts used for summary generation methods described in this paper.

For few-shot learning setting we randomly select samples from the training set. The example A.1 shows the prompt template we use for generating medical summaries in one-shot setting.

Example A.1. You are a helpful assistant. Your task is to summarize the given question based on the provided question and possibly helpful retrieved document. The retrieved document may or may not be useful for summarization.
Example:
QUESTION: {Example input text}
RETRIEVED DOCUMENT: {Example retrieved document}
SUMMARY: {Example summary}
QUESTION: {Test set input text}
RETRIEVED DOCUMENT: {Test set retrieved document}
SUMMARY:

The example A.2 shows the prompt template we use for generating medical summaries in two-shot setting.

Example A.2. You are a helpful assistant. Your task is to summarize the given question based on the provided question and possibly helpful retrieved document. The retrieved document may or may not be useful for summarization.
Examples:
QUESTION: {Example input text}
RETRIEVED DOCUMENT: {Example retrieved document}
SUMMARY: {Example summary}
QUESTION: {Example input text}
RETRIEVED DOCUMENT: {Example retrieved document}
SUMMARY: {Example summary}
QUESTION: {Test set input text}
RETRIEVED DOCUMENT: {Test set retrieved document}
SUMMARY: