RAG FOR ENHANCED MEDICAL QUESTION ANSWERING

Su Yang 223040097
MDS
Shenzhen, China
223040097@link.cuhk.edu.cn

ABSTRACT

Large language models (LLMs) exhibit strong general capabilities in natural language understanding but often struggle in specialized domains due to limited domain-specific knowledge. Traditional fine-tuning approaches to address this issue pose challenges such as overfitting and high computational costs. Retrieval-augmented generation (RAG) presents an efficient alternative by integrating external knowledge retrieval into the generation process, enhancing domain-specific performance without additional training. This report explores a RAG pipeline applied to the medical question answering (MedQA) task, evaluating various retrieval strategies and their influence on LLM effectiveness. The findings demonstrate that RAG significantly improves performance while emphasizing the critical role of knowledge base quality.

1 INTRODUCTION

Large language models (LLMs)Zhao et al. (2023)Vaswani et al. (2017), pretrained on billions of tokens, have demonstrated strong general capabilities in natural language understanding and manipulation. however, they often fall short in specialized domains due to a lack of expert-level knowledge. for example, in astrophysics, an LLM may fail to provide accurate answers if its training dataset contains limited relevant samples.

A common approach to address this limitation is to further train the LLM on curated, domain-specific datasets.Zhang et al. (2023) however, this method has two main drawbacks. first, the risk of model degradation—overfitting on a narrow dataset may improve performance on domain-specific tasks, but often at the expense of the model's general NLP capabilities. second, the training process demands substantial computational resources and energy.

To enhance an LLM's domain-specific performance without retraining, retrieval-augmented generation (RAG) offers a promising solutionFan et al. (2024)Gao et al. (2024). instead of updating model parameters, RAG retrieves relevant information from an external knowledge base and incorporates it into the prompt, enabling the LLM to respond with the necessary context. in this report, i present a RAG pipeline for the medical question answering task (MedQA), evaluating different RAG strategies and their impact on LLM performance. the experiments highlight RAG' s effectiveness in boosting performance without additional training, and underscore the importance of knowledge base quality in RAG systems.

2 RELATED WORK

The core of RAG centers on two aspects: modeling the knowledge base and the retrieval methodFan et al. (2024). Different modeling approaches lead to different RAG granularities, affecting retrieval efficiency and accuracy. Vector-based modeling is the most common, where texts, chunked or itemized, are embedded into vectors and retrieved via similarity metricsMa et al. (2023)Kim et al. (2023).

Graph-based modeling Sun et al. (2023), although more complex, enables graph-based inference with potentially stronger reasoning capabilities.

For retrieval, cosine similarity with prompt engineering is the most widely used method. However, it may struggle with complex queries requiring multi-hop reasoning. To address this, Sun et al. (2023) proposes retrieval based on graph inference, while Ma et al. (2023) trains a separate neural retriever.

In this report, to explore how knowledge granularity impacts RAG performance, I adopt vector-based knowledge modeling combined with vector-based retrieval.

3 METHODOLOGY

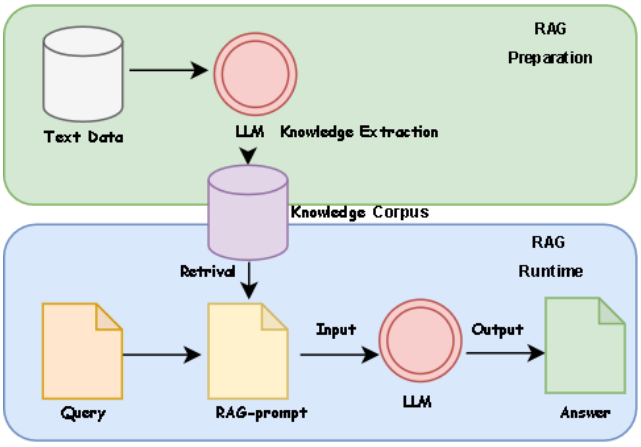

Figure 1: RAG pipeline

3.1 ONE-SHOT SIMILARITY MATCHING

The most straightforward approach to retrieval-augmented generation (RAG) for a medical question-answering task involves embedding all collected training question-answer (QA) samples during the data preparation phase. At inference time, the system retrieves the most semantically similar QA pair to the given test question and includes it as an example in the prompt—an approach known as one-shot promptingSahoo et al. (2024).

3.2 OPTION HINT SIMILARITY MATCHING

A more general approach to retrieval-augmented generation (RAG) in question-answering tasks involves constructing a knowledge base containing relevant information related to the question, rather than relying solely on example questions. For tasks that require broad knowledge coverage and multi-step reasoning, a graph structure is often employed to model entities and their relationships as implied in the knowledge text. While this solution ensures comprehensive knowledge representation for various question types, querying and performing inference over such a knowledge base can be resource-intensive, making it less suitable for real-world applications.

As a trade-off, *option hints* serve as neutral, concise descriptions of the key entities present in answer options, ensuring both precision and brevity. Large language models (LLMs) are employed to generate these hint texts for all entities in all options, based on the corresponding question context. A knowledge base is then constructed from these generated hints. During inference, RAG is applied only to the options: the most relevant hints are retrieved and appended to the question, providing additional context to assist the LLM in answering accurately.

An example of Option-hint generation

Prompt:

System: You are an expert in medical knowledge, now you need to do some tasks.Keep conciseness and preciseness.

Question:

A 39-year-old man presents to the emergency department because of progressively worsening chest pain and nausea that started at a local bar 30 minutes prior. The pain radiates to the epigastric area. He has a 5-year history of untreated hypertension. He has smoked 1 pack of cigarettes daily for the past 5 years and started abusing cocaine 2 weeks before his emergency room visit. The patient is diaphoretic and in marked distress. What should be the first step in management?

A. Diltiazem
B. Labetalol
C. Propranolol
D. Reassurance and continuous monitoring

Answer: Reassurance and continuous monitoring
For this QA example, you need to give neutral fact info for each option according to question context, so that later user can infer out the correct answer from these facts.

Expected Output:

1. Diltiazem: Calcium channel blocker; can reduce blood pressure but not first-line for cocaine-induced chest pain due to potential coronary vasospasm.
2. Labetalol: Combined alpha and beta blocker; may worsen coronary vasoconstriction in cocaine use.
3. Propranolol: Non-selective beta blocker; contraindicated in cocaine-induced symptoms due to unopposed alpha-adrenergic effects.
4. Reassurance and continuous monitoring: Initial step to assess for life-threatening conditions (e.g., myocardial infarction, aortic dissection) while avoiding harmful medications.

Figure 2: An example of Option-hint generation

4 EXPERIMENT SETTING

I use the MedQA dataset as the primary benchmark for the experiment. For option hint generation, the Deepseek-R1DeepSeek-AI et al. (2025) Chat model serves as the generator. In the one-shot similarity retrieval-augmented generation (RAG) setup, I employ MiniLM-L6-V2 from Sentence-Transformers as the encoder, and Faiss as the vector database for embedding storage and retrieval. It is worth noting that option hint retrieval is implemented using direct key hashing via a hashmap, rather than vector similarity.

An example of Option-hint prompting

Prompt:

System: You are an expert in doing QA exams. You need to answer some questions. You can refer to any examples, hints, and facts if provided.

Question:

Six days after undergoing surgical repair of a hip fracture, a previously healthy 79-year-old woman becomes agitated and confused. She is unarousable during the day but awake and impulsive at night, requiring frequent reorientation. Her husband reports that she usually drinks one to two glasses of wine per week. Her only current medication is oxycodone for pain. Vital signs are within normal limits. She appears distressed and is oriented to person but not to place or time. Neurologic examination reveals inattentiveness but no focal deficits. A urine dipstick test is normal.Which of the following is the most likely cause of her current condition?

A. Dementia
B. Opioid intoxication
C. Delirium
D. Urinary tract infection

Hint:

1. Dementia: Characterized by chronic, progressive cognitive decline. It typically does not present with sudden onset or fluctuating symptoms.
2. Opioid intoxication: Can cause sedation and confusion, but the symptoms are usually persistent rather than fluctuating.
3. Delirium: Presents with acute onset, fluctuating levels of consciousness, inattention, and disorientation. It is common in elderly patients following surgery.
4. Urinary tract infection: Can lead to delirium in older adults; however, a normal urine dipstick makes this diagnosis less likely.

Answer:

Expected Output:

C. Delirium.

Figure 3: An example of Option-hint prompting

5 Result Analysis

RAG-approach	Accuracy	Total Questions
None	0.85	100
One-shot	0.83	100
Option-hint	0.91	100

Table 1: Prompt CoT Report

The result shows that option-hint RAG greatly boost model performance with no need for training, which implies the importance of entity extraction and precise knowledge retrieval. One-shot similarity RAG performs worse than baseline, which is because it creates longer context, making it more difficult for semantic embedding, retrieval and LLM inference.

6 CONCLUSION

This study compares two retrieval-augmented generation (RAG) strategies for medical QA: one-shot similarity matching and option-hint similarity matching. Results on the MedQA dataset show that option-hint RAG achieves the highest accuracy (0.91), outperforming both the baseline (0.85) and one-shot RAG (0.83). The option-hint approach benefits from concise, entity-focused context and efficient key-based retrieval, avoiding the noise and length issues seen in one-shot QA pair retrieval. These findings highlight the effectiveness of targeted, structured augmentation in improving LLM performance without additional training.

DIVISION OF TASKS

The task is designed and implemented by myself only.

ACKNOWLEDGMENT

See details in https://nlp-course-cuhksz.github.io/.

REFERENCES

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z.F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J.L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R.J. Chen, R.L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S.S. Li, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. Submitted on 22 Jan 2025.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. A survey on rag meeting llms: Towards retrieval-augmented large language models. *arXiv preprint arXiv:2405.07123*, 2024. Submitted on 10 May 2024 (v1), last revised 17 Jun 2024 (this version, v3).

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2412.09956*, 2024. Submitted on 18 Dec 2023 (v1), last revised 27 Mar 2024 (this version, v5).

Gangwoo Kim, Sungdong Kim, Byeongguk Jeon, Joonsuk Park, and Jaewoo Kang. Tree of clarifications: Answering ambiguous questions with retrieval-augmented large language models. *arXiv preprint arXiv:2310.14696*, 2023. Submitted on 23 Oct 2023 (v1), accepted to EMNLP 2023.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. Query rewriting for retrieval-augmented large language models. *arXiv preprint arXiv:2305.14283*, 2023. URL https://doi.org/10.48550/arXiv.2305.14283. Submitted on 23 May 2023 (v1), last revised 23 Oct 2023 (this version, v3), EMNLP 2023.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*, 2024. Submitted on 5 Feb 2024 (v1), last revised 16 Mar 2025 (this version, v2).

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel M. Ni, Heung-Yeung Shum, and Jian Guo. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. *arXiv preprint arXiv:2307.07697*, 2023. URL https://doi.org/10.48550/arXiv.2307.07697. Submitted on 15 Jul 2023 (v1), last revised 24 Mar 2024 (this version, v6), Accepted by ICLR 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. URL https://doi.org/10.48550/arXiv.1706.03762. Submitted on 12 Jun 2017 (v1), last revised 2 Aug 2023 (this version, v7).

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*, 2023. Submitted on 21 Aug 2023 (v1), last revised 1 Dec 2024 (this version, v8).

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023. Submitted on 31 Mar 2023 (v1), last revised 11 Mar 2025 (this version, v16).