ReMAG-KR: Retrieval and Medically Assisted Generation with Knowledge Reduction for Medical Question Answering

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have significant potential for facilitating intelligent end-user applications in healthcare. However, hallucinations remain an inherent problem with LLMs, making it crucial to address this issue with extensive medical knowledge and data. In this work, we propose a Retrieve-and-Medically-Augmented-Generation with Knowledge Reduction (ReMAG-KR) pipeline, employing a carefully curated knowledge base using cross-encoder re-ranking strategies. The pipeline is tested on medical MCQ-based QA datasets as well as general QA datasets. It was observed that when the knowledge base is reduced, the model's performance decreases by 2-8%, while the inference time improves by 47%.

1 Introduction

Large Language Models (LLMs) like GPT-4 (Achiam et al., 2023), LLaMA-2 (Touvron et al., 2023a), and LLaMA-3 (Touvron et al., 2023b) have become highly efficient text generation tools with a significant variety of potential applications in a wide range of domains, like business, education, and healthcare. In healthcare, the potential to transform challenging tasks such as patient education (Jin et al., 2024), report generation (Shoham and Rappoport, 2023), and drug discovery (Kormilitzin et al., 2021; Unnikrishnan et al., 2023) is exemplary. This is primarily due to their ability to analyze large amounts of textual data and generate high-quality meaningful text as per end-user task requirements. However, there are specific challenges with deploying them in the healthcare industry. The shortage of medical data available for training/consumption by LLMs is one of the primary reasons for concern. A critical hurdle is the propensity of LLMs to produce false medical information (often termed hallucinations), misleading both patients and medical professionals. Additionally, in case of instructions that are too explicit or devoid of important details, LLMs fail to produce optimal results, which reduces their efficacy. LLMs may also reinforce biases learned from the training data, producing biased results towards particular groups of people based on constructs like gender, ethnicity, and socio-economic status.

For general tasks, the application of the concept of Retrieval-Augmented Generation (RAG) has shown promise. RAG systems incorporate external information retrieval into the LLM architecture. Previous research, such as Almanac (Zakka et al., 2024) and ChatENT (Long et al., 2023), has demonstrated improved LLM accuracy and reliability with this method. However, this kind of integration may also include unrelated or incorrect information, which could undermine the legitimacy and efficacy of the LLM. Including external knowledge sources raises issues with data consistency, privacy, security, and legal consequences. Furthermore, these methods frequently call for indexing and storing massive datasets, sometimes surpassing 200GB. Although RAG approaches perform excellently in general question-answering tasks, there is still uncertainty about their efficiency in healthcare. Regarding efficiency, retrievers trained on generic data often fall short of those optimized for particular domains (Li et al., 2022). This emphasizes the need for domain-specific training data, which can be costly and time-consuming to create, particularly in specialized fields like medicine. Moreover, conventional RAG techniques train the LLM and retriever separately (Steinberg et al., 2021; Agrawal et al., 2022), while other approaches include joint training of retrievers and LLMs (Wang et al., 2024). The retrieved informa-

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 140–145

tion and the LLM's capacity to process it for accurate output may generally lack semantic depth due to the nature of the training (Sarthi et al., 2024).

To address these challenges without additional computational costs, we propose a systematic approach that integrates RAG models built on a carefully curated knowledge base, with support for cross-encoder re-ranking strategies. First, keywords and entities from each query or question are extracted. Then, a web crawl is conducted to find each entity's top-15 relevant documents, which are used to build the knowledge base. Next, the retrieval based on the query and re-ranking of the results using MedCPT (Jin et al., 2023) is performed. Finally, responses are generated using LLMs, specifically LLaMa2 and LLaMa3. The rest of the article is structured as follows. Section 2 presents a detailed discussion on the proposed approach. Section 3 presents a discussion on the experiments performed and results observed, followed by conclusion and future work.

2 Methodology

Fig. 1 depicts the proposed approach consisting of five key phases – Keyword extraction, Document Retrieval, Knowledge base construction, Cross-encoder re-ranking, and response generation using LLMs. Given a set of medical questions Q, a set of keywords K_Q are extracted using KeyBERT (Grootendorst, 2020). For each keyword $k \in K_Q$, a ranked set of 15 relevant documents d_i are retrieved from the PubMed database, resulting in a comprehensive collection of medical documents D^* containing pertinent medical knowledge. Formally, for each question $q_i \in Q$, there exists a corresponding correct answer a_i^* within a set of options A_i, such that $a_i^* \in A_i$. The model M utilizes the query q and relevant document d to produce a predicted answer (as per Eq. (1), where, $d \in D^*$ and d_R is the retrieved document based on the query, and Eq. (2) where θ represents the model's parameters).

$$Document\ Retrieval\ d_R = p(d \mid q) \quad (1)$$

$$Answer\ Prediction\ a = p(a \mid q, d_R, \theta) \quad (2)$$

In the Keyword Extraction phase, at least three keywords are extracted from all queries. For each query Q, KeyBERT is used to extract at least three related medical keywords or key phrases K_Q. This

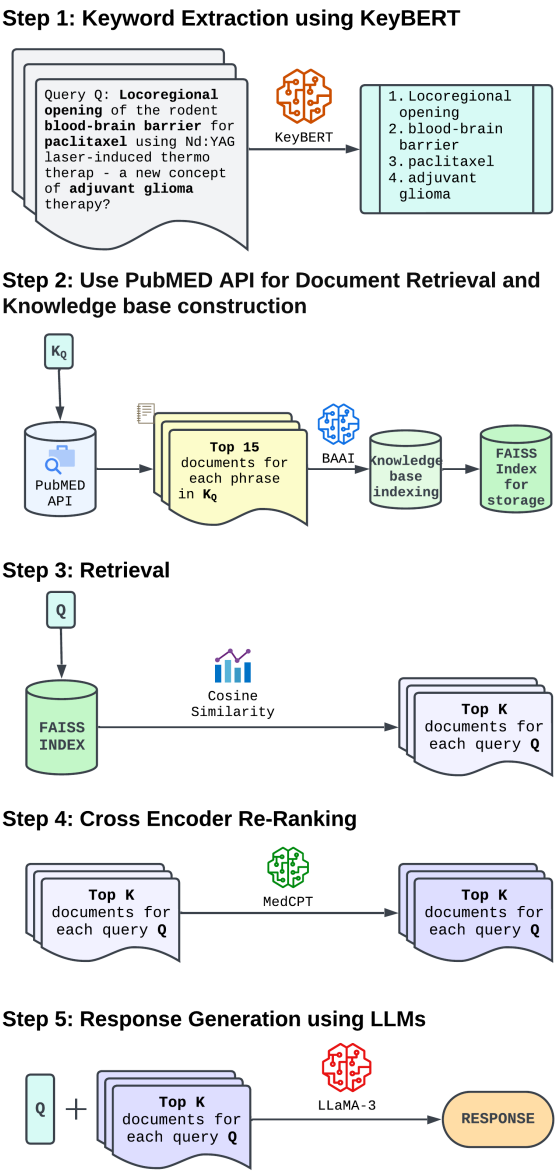

Figure 1: Proposed ReMAG-KR Framework

ensures that individual words and phrases relevant to the medical context are captured accurately. Following this, in the Knowledge Base Indexing and Storage phase, the PubMed API is employed to retrieve 15 relevant articles for each identified keyword or keyphrase. This retrieval process results in a substantial corpus of about 600,000 articles, providing a focused subset of the extensive PubMed database (Canese and Weis, 2013), consisting of 24.9 million articles. The collected articles are then transformed into embedding vectors through the BAAI embed-

ding model (Zhang et al., 2024). This model converts text data into a format that can be efficiently processed for similarity searches. Finally, the FAISS (Facebook AI Similarity Search) VectorStore index is used to store these embedding vectors. FAISS is optimized for high-speed similarity searches on large datasets, making it suitable for handling the extensive medical corpus generated.

For facilitating retrieval, MedCPT was used in the process of document retrieval. For this, we compute the cosine similarity score between the query embedding q and each document embedding d_i. The top k documents were chosen based on query similarity. MedCPT was utilized to streamline the retrieval process and ensure that the most relevant documents were retrieved. For this, the cosine similarity score was computed between the query embedding q and the embedding of each document d_i. The cos-sim score measures how similar two vectors are in orientation and magnitude, with a higher score indicating greater similarity. By calculating this score for each document in the database, the system can effectively rank them based on their relevance to the query. Using the computed scores, the top k documents were selected for retrieval, ensuring that the retrieved documents are closely aligned with the user's query.

The cross-encoder re-ranker MedCPT is advantageous in re-ranking the top k extracted articles for generating the top n articles (where $n = k$), enhancing the relevance of the information produced. MedCPT was chosen as retriever and re-ranker due to its First-stage dense retriever (MedCPT retriever) and the Second-stage re-ranker (MedCPT re-ranker). The MedCPT retriever contains a query encoder (QEnc) and an article encoder (DEnc), both initialized by PubMedBERT. It is trained on 255M query-article pairs from PubMed search logs and in-batch negatives. The MedCPT re-ranker is a transformer cross-encoder (CrossEnc) initialized by PubMedBERT. It is trained on 18M semantic query-article pairs and localized negatives derived from the pre-trained MedCPT retriever.

Upon re-ranking retrieved articles, they are merged with the original query and provided as input to the LLM, which generates a response, represented as a, based on the amalgamation of the query and the re-ranked articles. The generated response is then evaluated by comparing it with the ground-truth answers, serving as a metric for assessing the performance of the LLM in understanding and responding to the given query. We have utilized two specific LLMs for our experiments, namely LLaMA-2 and LLaMA-3. These models have been selected based on their capabilities and suitability for the task at hand. We aim to evaluate these LLMs' effectiveness in generating accurate responses when presented with queries and relevant document contexts.

3 Experiments and Results

Experiments were conducted on the benchmark MIRAGE dataset (Xiong et al., 2024) for the multiple choice questions-based QA tasks. This included 7,663 questions from five commonly used QA datasets in biomedicine (MMLU-Med, MedQA-US, MedMCQA, PubMedQA (Jin et al., 2019), BioASQ (Y/N)) (Tsatsaronis et al., 2015). For Subjective QA task, the datasets LiveQA (Abacha et al., 2017) and ExpertQA-Med (Malaviya et al., 2023) were chosen, with 3,479 subjective questions and answers. Standard metrics like accuracy, precision, recall, and F1-score were used for the evaluation. The generated text quality and relevancy were assessed using BLEURT, BERTScore, MoverScore, and ROUGE-L. For MCQ-based QA tasks, the MED-RAG model was used as the baseline, while KG-Rank (Yang et al., 2024) was considered for subjective tasks due to its novelty and outstanding scores.

3.1 Results and Discussion

MCQ-based tasks: Table 1 shows the results for this task, and it is evident that the proposed ReMAG-KR underperformed on datasets including MMLU-Med, MedQA-US, MedMCQA, PubMedQA, and BioASQ-Y/N. Compared to MEDRAG's 73.09 average accuracy, our approach produced 66.32. Likewise, our approach averaged 58.74 for the F1 score, whereas MEDRAG scored 66.69. Despite the lag in performance, the proposed ReMAG-KR showed a notable efficiency advantage, as seen in Table 3. The inference time was nearly one-third that of MEDRAG, primarily because our knowledge base contains only 600,000 documents compared to MEDRAG's extensive 25 million corpus.

Subjective tasks: Using LLaMA-2 and LLaMA-3 models, we analyzed the ExpertQA-Med and

Table 1: k-sample average performance comparison between MED-RAG and ReMAG-KR (proposed) for MIRAGE benchmark.

Dataset	Method	**LLaMA-2** (k = 200)				**LLaMA-3** (k = 200)			
		Accuracy	**Precision**	**Recall**	**F1**	**Accuracy**	**Precision**	**Recall**	**F1**
MMLU-Med	MEDRAG	73.35	73.01	74.02	75.84	77.73	78.93	75.08	76.47
	ReMAG-KR	66.32	67.43	70.32	68.34	71.72	72.23	74.15	70.38
MedQA-US	MEDRAG	66.72	65.72	66.15	64.53	70.26	68.38	69.15	64.53
	ReMAG-KR	58.74	58.77	59.77	55.18	65.84	64.97	61.77	60.18
MedMCQA	MEDRAG	54.94	56.10	56.50	55.98	60.86	61.89	60.50	60.98
	ReMAG-KR	50.38	52.12	53.11	52.47	56.40	55.50	57.11	56.47
PubMedQA	MEDRAG	66.52	63.56	64.96	65.30	69.30	71.81	69.47	68.38
	ReMAG-KR	58.92	60.46	60.47	60.87	63.19	62.50	63.23	62.60
BioASQ-Y/N	MEDRAG	85.05	83.56	85.96	85.30	89.30	87.81	87.47	88.38
	ReMAG-KR	79.72	80.46	80.47	80.87	84.19	84.50	83.23	83.60

Table 2: Performance comparison for LiveQA and ExpertQA-Med between KGRank and ReMAG-KR (proposed)

Dataset	Method	**LLaMA-2**				**LLaMA-3**			
		ROUGE-L	**BERTScore**	**MoverScore**	**BLEURT**	**ROUGE-L**	**BERTScore**	**MoverScore**	**BLEURT**
ExpertQA-Med	KGRank	28.02	86.01	57.02	47.14	29.03	86.93	58.08	48.47
	ReMAG-KR	28.32	82.43	53.32	48.34	28.72	84.23	57.15	47.38
LiveQA	KG-Rank	19.72	82.02	54.15	40.34	20.26	83.38	54.65	40.53
	ReMAG-KR	17.84	79.77	55.77	39.18	18.84	81.97	54.77	41.18

Table 3: Inference time for MED-RAG and ReMAG-KR.

Dataset	ReMAG-KR (in s)	MedRAG (in s)
MMLU-Med	**680**	1380
MedQA-US	**720**	1500
MedMCQA	**970**	1432
PubMedQA	**703**	1290
BioASQ-Y/N	**400**	1000
AVG	**694.6**	1320.4

LiveQA datasets and compared the effectiveness of the RR and ReMAG-KR approaches (Refer Table 2. With LLaMA-2 and LLaMA-3, respectively, KG-Rank obtained a ROUGE-L of 28.02 and 29.03 and a BERTScore of 86.01 and 86.93 for ExpertQA-Med. ReMAG-KR obtained BERTScore and ROUGE-L scores of 82.43, 84.23, 28.32, and 28.72, respectively. For LiveQA, KG-Rank obtained BERTScores of 82.02 and 83.38 in addition to ROUGE-L scores of 19.72 and 20.26. For BERTScore, ReMAG-KR scored 79.77 and 81.97, and for ROUGE-L, 17.84 and 18.84. Both techniques showed comparable performance with equal inference times.

4 Conclusion and Future Work

An approach for LLM-based retrieval and medically assisted generation with a tactically reduced knowledge base was presented in this article. Experiments revealed that our approach reduces inference time by 47% with a small compromise in performance (of around 2-8%). Performance for subjective QA tasks was also comparable with the state-of-the-art approaches in this field. We plan to extend the proposed approach by enriching the quality of retrieved documents, while maintaining a reduced inference time. Simple, vanilla RAG-based approaches fail to capture semantically deep information hidden within medical text, thus, the use of a single corpus for generating a knowledge base encompassing multiple sources of information could be attempted. We also plan to introduce two other components into the RAG pipeline – multi-hop question answering and question decomposition. This involves breaking down a complex query into sub-queries and enriching the quality of retrieved documents. Adopting domain-specific models like PMC and MedLLaMA may further boost the model's ability to handle the intricacies and nuances inherent in medical data.

References

Asma Ben Abacha, Eugene Agichtein, Yuval Pinter, and Dina Demner-Fushman. 2017. Overview of the medical question answering task at trec 2017 liveqa. In *TREC*, pages 1–12.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. *arXiv preprint arXiv:2205.12689*.

Kathi Canese and Sarah Weis. 2013. Pubmed: the bibliographic database. *The NCBI handbook*, 2(1).

Maarten Grootendorst. 2020. Keybert: Minimal keyword extraction with bert. https://github.com/MaartenGr/KeyBERT.

Mingyu Jin, Qinkai Yu, Chong Zhang, Dong Shu, Suiyuan Zhu, Mengnan Du, Yongfeng Zhang, and Yanda Meng. 2024. Health-llm: Personalized retrieval-augmented disease prediction model. *arXiv preprint arXiv:2402.00746*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651.

Andrey Kormilitzin, Nemanja Vaci, Qiang Liu, and Alejo Nevado-Holgado. 2021. Med7: A transferable clinical natural language processing model for electronic health records. *Artificial Intelligence in Medicine*, 118:102086.

Huayang Li, Yixuan Su, Deng Cai, Yan Wang, and Lemao Liu. 2022. A survey on retrieval-augmented text generation. *arXiv preprint arXiv:2202.01110*.

Cai Long, Deepak Subburam, Kayle Lowe, André dos Santos, Jessica Zhang, Sang Hwang, Neil Saduka, Yoav Horev, Tao Su, David Cote, et al. 2023. Chatent: Augmented large language model for expert knowledge retrieval in otolaryngology-head and neck surgery. *medRxiv*, pages 2023–08.

Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. 2023. Expertqa: Expert-curated questions and attributed answers. *arXiv preprint arXiv:2309.07852*.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D Manning. 2024. Raptor: Recursive abstractive processing for tree-organized retrieval. *arXiv preprint arXiv:2401.18059*.

Ofir Ben Shoham and Nadav Rappoport. 2023. Cpllm: Clinical prediction with large language models. *arXiv preprint arXiv:2309.11295*.

Ethan Steinberg, Ken Jung, Jason A Fries, Conor K Corbin, Stephen R Pfohl, and Nigam H Shah. 2021. Language models are an effective representation learning technique for electronic health record data. *Journal of biomedical informatics*, 113:103637.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16:1–28.

Reshma Unnikrishnan, Sowmya Kamath, and VS Ananthanarayana. 2023. Efficient parameter tuning of neural foundation models for drug perspective prediction from unstructured socio-medical data. *Engineering Applications of Artificial Intelligence*, 123:106214.

Junda Wang, Zhichao Yang, Zonghai Yao, and Hong Yu. 2024. Jmlr: Joint medical llm and retrieval training for enhancing reasoning and professional question answering capability. *arXiv preprint arXiv:2402.17887*.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. *arXiv preprint arXiv:2402.13178*.

Rui Yang, Haoran Liu, Qingcheng Zeng, Yu He Ke, Wanxin Li, Lechao Cheng, Qingyu Chen, James Caverlee, Yutaka Matsuo, and Irene Li. 2024. Kg-rank: Enhancing large language models for medical qa with

knowledge graphs and ranking techniques. *arXiv preprint arXiv:2403.05881*.

Cyril Zakka, Rohan Shad, Akash Chaurasia, Alex R Dalal, Jennifer L Kim, Michael Moor, Robyn Fong, Curran Phillips, Kevin Alexander, Euan Ashley, et al. 2024. Almanac—retrieval-augmented language models for clinical medicine. *NEJM AI*, 1(2):AIoa2300068.

Mingtian Zhang, Shawn Lan, Peter Hayes, and David Barber. 2024. Mafin: Enhancing black-box embeddings with model augmented fine-tuning. *arXiv preprint arXiv:2402.12177*.