Augmenting Black-box LLMs with Medical Textbooks for Biomedical Question Answering

Yubo Wang, Xueguang Ma, Wenhu Chen*
University of Waterloo
{y726wang, x93ma, wenhuchen}@uwaterloo.ca†

Abstract

Large Language Models (LLMs) like ChatGPT have demonstrated impressive abilities in generating responses based on human instructions. However, their use in the medical field can be challenging due to their lack of specific, in-depth knowledge. In this study, we present a system called LLMs Augmented with Medical Textbooks (LLM-AMT) designed to enhance the proficiency of LLMs in specialized domains. LLM-AMT integrates authoritative medical textbooks into the LLMs' framework using plug-and-play modules. These modules include a *Query Augmenter*, a *Hybrid Textbook Retriever*, and a *Knowledge Self-Refiner*. Together, they incorporate authoritative medical knowledge. Additionally, an *LLM Reader* aids in contextual understanding. Our experimental results on three medical QA tasks demonstrate that LLM-AMT significantly improves response quality, with accuracy gains ranging from 11.6% to 16.6%. Notably, with GPT-4-Turbo as the base model, LLM-AMT outperforms the specialized Med-PaLM 2 model pre-trained on a massive amount of medical corpus by 2-3%. We found that despite being 100× smaller in size, medical textbooks as a retrieval corpus are proven to be a more effective knowledge database than Wikipedia in the medical domain, boosting performance by 7.8%-13.7%.

1 Introduction

Recent advancements in Large Language Models (LLMs) have opened new possibilities in AI for the medical domain, enabling them to comprehend and communicate through language. The promise of these models is underscored by their performance on medical question-answering datasets (Zhang et al., 2018; Pal et al., 2022; Jin et al., 2019).

LLMs are typically trained to encode world knowledge in their parameters. However, this can lead to information loss and "memory distortion" (Peng et al., 2023), resulting in the generation of plausible but incorrect content. Augmenting LLMs with external knowledge has become an interest to mitigate this, but fine-tuning LLM parameters for this purpose is often costly, especially as model sizes increase (Luo et al., 2022; Gao et al., 2022a; Singhal et al., 2023).

The Retrieval-Augmented Generation (RAG) framework provides an efficient solution to the limitations of fine-tuning in open-domain QA, pairing a retriever for sourcing relevant documents with a reader for answer extraction (Lewis et al., 2020; Karpukhin et al., 2020; Izacard et al., 2022). Enhancements in retrieval accuracy (Wu et al., 2021; Izacard et al., 2021) and reader model co-training (Lewis et al., 2020; Izacard et al., 2022) have been made, with current iterations leveraging LLMs as readers to adapt specifically to their capabilities (Shi et al., 2023b). However, many rely on general knowledge bases like Wikipedia or search engines such as Google and Bing. Such sources, while vast, might lack depth in domain-specific areas like medical or financial fields. Tapping into specialized resources, such as authoritative textbooks, could yield deeper insights into complex domains.

The effectiveness of the retrieval process in enhancing LLMs with additional information is heavily reliant on the quality of retrieval. If the retrieval process is inaccurate or contains misinformation, the utility of the RAG process can be significantly influenced (Li et al., 2022; Tan et al., 2022; Shi et al., 2023a). To address these challenges, several approaches have been proposed, such as HyDE (Gao et al., 2022b) and query2doc (Wang et al., 2023), which aim to improve retrieval by generating hypothetical documents to expand the query. On the other hand, methods like self-RAG (Asai et al., 2023) have introduced retrieval results reflection to filter the retrieved information for better generation. Building on these advancements, we have

*Corresponding author
†Code is available at: https://github.com/TIGER-AI-Lab/LLM-AMT

Findings of the Association for Computational Linguistics: EACL 2024, pages 1754–1770

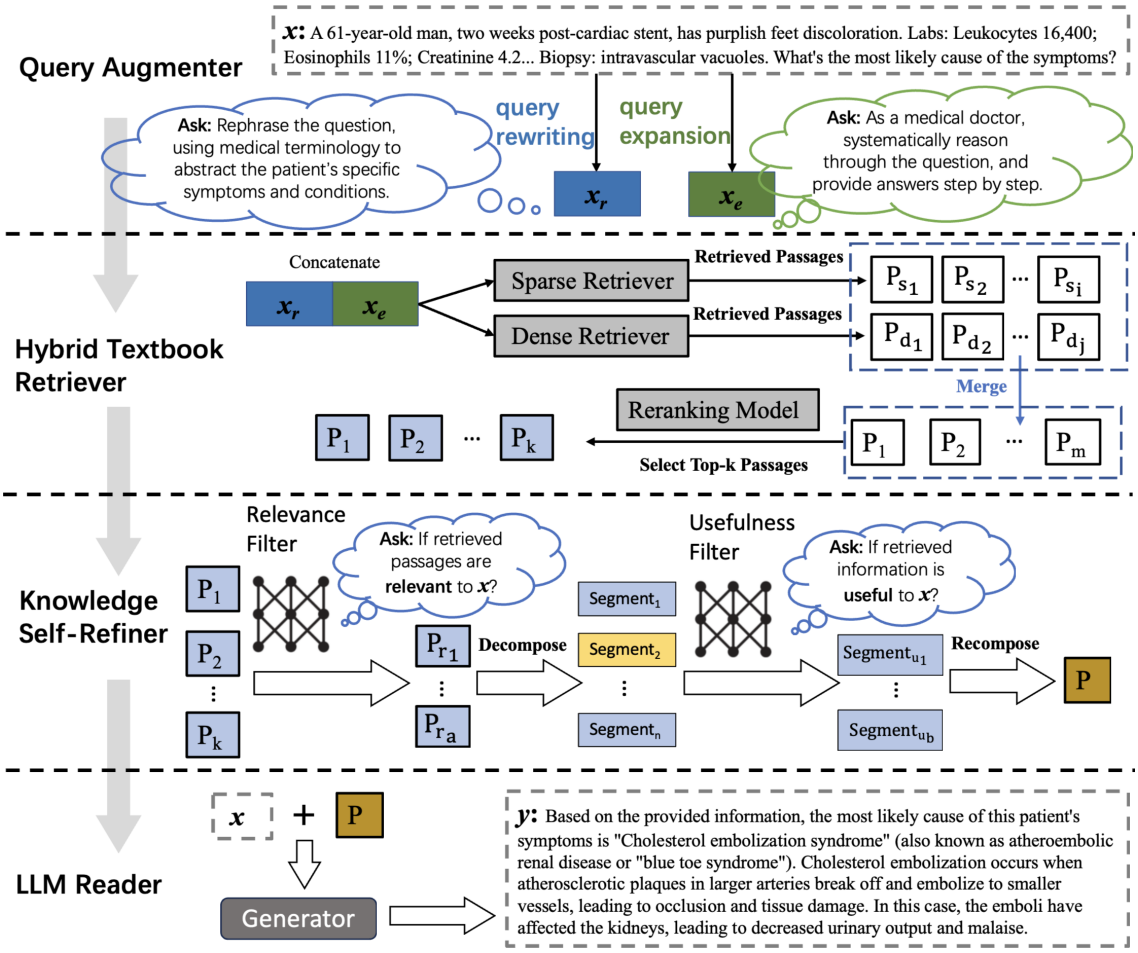

Figure 1: Overview of our proposed pipeline. GPT-3.5-Turbo is prompted to perform a series of tasks including query rewriting, query expansion, relevance evaluation, and usefulness evaluation. Details of the case presented in the figure can be found in Appendix A.7.

developed Query Augmenter and Knowledge Self-Refiner specifically tailored for the medical RAG paradigm. Combining with Hybrid Textbook Retriever, we have created a novel framework, LLM-AMT, specifically designed for the medical domain. This framework leverages medical textbooks to provide LLMs with high-quality, reliable knowledge. This approach ensures that the information used during the generation process is not only accurate but also highly relevant to answer the questions.

Our evaluations across MedQA-USMLE, MedQA-MCMLE, and MedMCQA datasets demonstrate LLM-AMT's outperformance over GPT-3.5, achieving 11.6% to 16.6% higher accuracy. Notably, our approach leverages the in-depth knowledge from textbooks, eclipsing Wikipedia's broader scope and a 7.8% to 13.7% accuracy gain. With GPT-4-Turbo as a base, LLM-AMT further exceeds the medically pre-trained Med-PaLM 2 by 2.3% to 2.7%. Additionally, human evaluations reveal a significant 16% reduction in hallucination occurrences during open-ended QA tasks, showcasing the model's improved reliability.

Our contributions are fourfold:

1. We propose LLM-AMT, an LLM pipeline augmented with medical textbooks, which enhances model accuracy and domain-specific expertise.
2. We introduce the Knowledge Self-Refiner, a novel component that implements self-refinement mechanisms for RAG models within the medical domain.
3. We demonstrate the significant impact of domain-specific textbooks on LLM performance through comprehensive experiments, opening new avenues for research in specialized knowledge integration.
4. We conduct an ablation study analyzing the roles of key components in our medical pipeline.

2 LLM-AMT

In this paper, we introduce LLM-AMT, a dedicated process tailored for answering biomedical questions. Figure 1 provides an overview of the process pipeline, which consists of four main components. The **Query Augmenter** rewrites and expands the input question $\boldsymbol{x}$ into a rewritten version $\boldsymbol{x}_r$ and an expanded version $\boldsymbol{x}_e$. Following this, the **Textbook Retriever** collects related passages $P_1, P_2, \ldots, P_k$ from textbooks by concatenating the augmented queries $\boldsymbol{x}_r$ and $\boldsymbol{x}_e$. Next, the **Knowledge Self-Refiner** employs a relevance filter to remove non-pertinent passages and then applies a usefulness filter to discard unhelpful segments from the remaining passages, producing refined knowledge P. Finally, the **LLM Reader** utilizes this refined knowledge to construct the final answer.

2.1 Query Augmenter

To address the challenges in biomedical question answering, where non-standard terms and discrete numerical values often impede effective information retrieval, we introduce a query augmenting module tailored for the medical domain. Our augmenter enhances queries by transforming ambiguous language and integrating key medical terms, which are crucial for accurate retrieval.

The module consists of two principal components: **query rewriting** and **query expansion**. The motivation behind query rewriting is to map colloquial or non-standard expressions to standardized medical terminology. For instance, our system converts phrases such as "high blood cell count" to the precise term "leukocytosis." Such transformation is pivotal because it aligns patient-described symptoms with professional language. This method ensures the retention of crucial information and translates it into the language of medicine, thus making the query more suitable for professional databases.

On the other hand, query expansion leverages the LLM's ability to reason through problems without external evidence, invoking a chain-of-thought approach. By instructing the LLM with *"As a medical doctor, systematically reason through the question, and provide answers step by step."* We introduce additional relevant medical terms into the query. This preemptive reasoning extracts more directions for retrieval and enhances the likelihood of accessing pertinent information.

Metric	Textbooks	Wikipedia
# of paragraphs	347,797	21,015,324
# of tokens	27,458,075	2,162,169,361

Table 1: Overall statistics of the document collection in textbooks and Wikipedia. The Wikipedia dump is from the DPR work (Karpukhin et al., 2020), where Wikipedia documents are split into 100-word units.

2.2 Textbook Retrieval Corpus

Medical textbooks, as the epitome of knowledge in human medicine, serve as an invaluable external knowledge source. While knowledge bases like Wikipedia provide general information, textbooks offer richer and more specialized domain knowledge. In contrast to search engines like Google Search or Bing Search, the information in textbooks is more reliable. Furthermore, textbooks offer clear and concise information, making them a reliable source for text-based retrieval. For our study, we eliminated irrelevant information, such as diagrams and references, to ensure a focused, text-centric corpus. Additionally, longer paragraphs in the textbook were broken down according to periods to obtain the smallest unit for retrieval, making it easier for the LLM reader to use them as context for questions. In this paper, we utilized 51 textbooks from the MedQA dataset (Jin et al., 2021), which are designated as the official preparation materials for the medical licensing exams.

An overview of the statistics for the document collection in both the textbooks and Wikipedia can be seen in Table 1. Our textbook corpus is substantially smaller in scale than Wikipedia. While Wikipedia comprises millions of paragraphs and billions of tokens, the textbook corpus, though specialized, contains fewer than 350,000 paragraphs and just over 27 million tokens. This size difference emphasizes the textbooks' concentrated domain-specific knowledge.

2.3 HybTextR (Hybrid Textbook Retriever)

We integrate various types of retrievers in our textbook retrieval module to optimize performance, which we refer to as the HybTextR. For **sparse retrieval**, we follow the SPLADE (Formal et al., 2021) method. The query and document are encoded separately by BERT, and the MLM layer representation (with dimension 30k) for each token is aggregated at the maximum as the text representation. The ReLU function is used to truncate the

weights in the representation to be non-negative so that it can fit into an inverted index after quantization at search time. The sparsity of this representation is effectively managed by a FLOP loss during the training stage. For **dense retrieval**, we follow the standard pipeline proposed in the DPR work, where the query and document embeddings are taken from the CLS token's dense representation in the last layer output (with dimension 768). At search time, a k-NN search is conducted to retrieve the top relevant passages for the given query. For **reranking**, our model is a BERT-based cross-encoder. In line with the approach described in Tevatron by Gao et al. (Gao et al., 2022c), we concatenate the query and the retrieved passages using the [SEP] token and employ the representation from the [CLS] token to predict relevance scores. This cross-encoder setup allows for a more nuanced understanding of the relationship between the query and the passages, leading to improved ranking accuracy.

A core problem in the task is how to create supervised data for the neural retriever. As there is no human relevance judgment for the passages, we treat "helpful" passages as positive passages. We first identify questions that GPT-3.5-Turbo answers incorrectly when provided without any contextual evidence. Then, using BM25, we recall n passages, where $n = 32$, and concatenate each of them with the original question to serve as its context. Subsequently, GPT-3.5-Turbo is prompted to answer this question. Passages resulting in correct answers are treated as positive samples, whereas those leading to incorrect answers are categorized as hard negatives. Additionally, a subset of passages is randomly chosen to act as easy negative samples.

In the full pipeline of our knowledge retrieval stage, we utilize a fusion of sparse retrieval and dense retrieval as the first-phase recall model. Specifically, we merge and deduplicate passages returned by the sparse retriever (from P_{s_1} to P_{s_i}) and those returned by the dense retriever (from P_{d_1} to P_{d_j}), resulting in a total of m unique passages as illustrated by Figure 1. These passages are then reordered by the reranker. Finally, the top-k passages are selected for further processing. In this study, i, j, and k are all set to a fixed value of 32.

2.4 Knowledge Self-Refiner

The motivation for introducing the Knowledge Self-Refiner is driven by the structure of medical textbooks, which typically present elongated passages dense with information on a particular topic. Given that not all content within these passages is pertinent to addressing specific questions, and unfiltered content may lead to a diffusion of the LLM Reader's focus, we implemented the Knowledge Self-Refiner to streamline the information.

Given the retrieved passages $P_1, P_2, \ldots, P_k$, our Knowledge Self-Refiner begins by applying a relevance filter to exclude off-topic passages. This filter performs a binary classification on each passage to determine its relevance, resulting in a subset $P_{r_1}, P_{r_2}, \ldots, P_{r_a}$. Subsequent to this initial filtering, a decompose-then-recompose algorithm is employed. Passages are segmented at sentence boundaries according to heuristic rules. Segments that do not meet a minimum length threshold are merged with adjacent segments to ensure the combined length does not exceed 80 words. This process yields segments $segment_1, segment_2, \ldots, segment_n$, each balancing substance with brevity.

These segments are then passed through a stringent usefulness filter, which performs a binary classification to assess the utility of each segment. This filter distills the content to retain only the segments $segment_{u_1}, segment_{u_2}, \ldots, segment_{u_b}$ that are deemed most useful for the LLM to generate accurate responses. The final refined knowledge set, denoted by **P**, is composed of these selected segments, offering a concentrated and relevant reservoir of information for the LLM.

This two-tiered filtering approach is specifically designed to address the high density of knowledge points in medical texts, which typically feature lengthy passages with only a few sentences of critical importance. Applying both filters at the passage level could lead to the inclusion of extensive but irrelevant content, potentially diverting the attention of the LLM Reader in the subsequent Retrieval-Augmented Generation (RAG) system. For an in-depth discussion, see Appendix A.7.

3 Experiments

3.1 Datasets

We evaluate LLM-AMT on three medical open-domain multiple-choice QA datasets as follows:

MedQA-USMLE and MedQA-MCMLE (Jin et al., 2021) originate from professional medical board exams in the USA and Mainland China, where doctors are evaluated on their professional knowledge and ability to make clinical decisions.

Question #	MedQA-USMLE	MedQA-MCMLE	Med-MCQA
Train	10,178	27,400	182,822
Dev	1,272	3,425	4,183
Test	1,273	3,426	6,150

Table 2: Number of Questions in MedQA-USMLE, MedQA-MCMLE, and MedMCQA

In addition to the questions and corresponding answers, the datasets also provide associated medical textbook materials. For the USMLE, the MedQA-USMLE dataset includes text extracted from a total of 18 English medical textbooks used by USMLE candidates. For the MCMLE, the MedQA-MCMLE dataset features materials from 33 simplified Chinese medical textbooks. These are designated as the official textbooks for preparing for the medical licensing exam in Mainland China.

MedMCQA (Pal et al., 2022) encompasses a broad spectrum of 2,400 healthcare topics and 21 distinct medical subjects. The diversity of questions contained within MedMCQA illustrates the challenges that are unique to this dataset. As the questions are derived from both real-world scenarios and simulated examinations, they are meticulously crafted by human experts in the field. Consequently, these questions could serve as a comprehensive evaluation of a medical practitioner's professional competencies and expertise.

Table 2 shows the detail of train/dev/test splits of the datasets. We evaluate our pipeline and conduct ablation studies on the test sets of each dataset.

3.2 Baselines

Our evaluations encompass two primary categories of models. The first group consists of the *Closed-Book Models*, which are pre-trained or fine-tuned specifically for the medical domain. These models rely on their internal knowledge and do not access external databases or texts during the question-answering process. Notable models in this category include **BioBERT**, **SciBERT**, **BioLinkBERT**, **PubmedBERT**, **Flan-PaLM (540B)**, **Meditron-70B**, **Med PaLM 2** (Lee et al., 2020; Beltagy et al., 2019; Yasunaga et al., 2022; Gu et al., 2021; Singhal et al., 2022; Chen et al., 2023; Singhal et al., 2023). It is important to note that data marked with an asterisk* were obtained directly from the respective authors' published works.

The second group, *Wikipedia-Augmented Models*, leverages the knowledge embedded in Wikipedia to assist in the medical QA task. Key models in this category are **Variational ODQA** (Liévin et al., 2023), **Codex 5-shot CoT** (Liévin et al., 2022). We have separately employed **LLaMA-2-13B**, **GPT-3.5-Turbo**[1] and **GPT-4-Turbo** as readers, enhanced by the knowledge retrieved from Wikipedia to answer questions.

3.3 Implementation Details

We employ OpenAI's GPT-3.5-Turbo as our LLM readers in different experiments. LLaMA-2-13B and GPT-4-Turbo are only used in the main result experiments of Table 3. Subsequent Ablation Studies only utilize GPT-3.5 as the generator. GPT-3.5-Turbo, accessed via its API[2], handled query rewriting during the augmentation phase. In the evidence retrieval stage, SPLADE acts as our sparse retriever, DPR is the dense retriever, and we incorporate a cross-encoder for reranking. The MS-MARCO dataset (Nguyen et al., 2016) is our primary training source for our zero-shot model. Specifics related to fine-tuning, such as batch size, learning rate, and training rounds, can be found in the supplementary material.

3.4 Main Result

In Table 3, we compare various state-of-the-art models with our proposed pipeline on MedQA and MedMCQA datasets.

Our experiments reveal that incorporating textbook knowledge with our proposed method significantly enhances the performance of GPT-3.5-Turbo and GPT-4-Turbo when compared to closed-book models. While Wikipedia is a rich information source, its content may be too generalized and often lacks the necessary depth for specialized fields such as medicine. Therefore, the smaller performance gains observed when utilizing Wikipedia as an external knowledge base may be due to the fact that these large language models have already incorporated Wikipedia data during pre-training. To explore the effectiveness of Wikipedia as a retrieval corpus, we employed two distinct retrievers: a publicly available pre-finetuned DPR from other researchers (Karpukhin et al., 2020), and our own fine-tuned HybTextR system using the Wikipedia corpus as training data. Both methods indicated that a textbook corpus is more useful compared to

[1]In this paper, "GPT-3.5" denotes *gpt-3.5-turbo-0125*. Similarly, references to "GPT-4" imply *gpt-4-0125-preview.*.

[2]https://platform.openai.com/docs/guides/gpt

Method	Retriever	MedQA-USMLE	MedMCQA
Closed-Book Model			
Random	-	20.0	25.0
BioBERT*	-	36.7	37.0
SciBERT*	-	-	39.0
BioLinkBERT*	-	45.1	-
PubmedBERT*	-	50.3	41.0
LLaMA	-	31.4	35.7
GPT-3.5	-	51.3	53.9
Flan-PaLM (540B)*	-	67.6	-
Meditron-70B*	-	70.2	-
GPT-4	-	81.7	70.5
Med-PaLM 2*	-	85.4	72.3
Wikipedia-Augmented Model			
Variational ODQA*	BM25+DPR	55.0	62.9
Codex 5-shot CoT*	BM25	60.2	62.7
LLaMA + Wikipedia	DPR	38.6	40.5
LLaMA + Wikipedia	HybTextR	39.9	41.3
GPT-3.5 + Wikipedia	DPR	52.8	56.8
GPT-3.5 + Wikipedia	HybTextR	54.2	57.7
GPT-4 + Wikipedia	DPR	80.6	69.8
GPT-4 + Wikipedia	HybTextR	81.5	71.2
Textbook-Augmented Model			
LLM-AMT (LLaMA)	HybTextR	42.2	43.8
LLM-AMT (GPT-3.5)	HybTextR	67.9	65.5
LLM-AMT (GPT-4)	HybTextR	**88.1**	**74.6**

Table 3: Performance of various state-of-the-art models on MedQA and MedMCQA datasets.

Method	MedQA-USMLE	MedQA-MCMLE	Med-MCQA
GPT-3.5-Turbo	51.3	58.2	53.9
+ retriever	58.6	61.2	57.1
+ retriever + augmented query	62.0	65.4	63.1
+ retriever + knowledge self-refiner	63.9	68.1	64.4
+ retriever + augmented query + knowledge self-refiner	65.0	68.8	65.1
+ finetuned retriever	61.2	62.3	58.7
+ finetuned retriever + augmented query	64.1	68.9	63.4
+ finetuned retriever + knowledge self-refiner	65.7	70.3	64.8
+ finetuned retriever + augmented query + knowledge self-refiner	**67.9**	**72.6**	**65.5**

Table 4: Performance comparison (% accuracy) of various approaches on three medical QA datasets. The table showcases the incremental improvements gained by integrating different components. Specifically, the retriever employed is HybTextR, and the LLM Reader is GPT-3.5-Turbo.

Wikipedia for enhancing medical QA performance. This is evidenced by a 13.7% increase over the *GPT-3.5 + Wiki* for MedQA and a 7.8% increase for MedMCQA, highlighting the significance of integrating deep, specialized medical knowledge overbroad, surface-level information sources.

Moreover, when leveraging the more sophisticated GPT-4 as the base model, our approach surpasses the performance of specialized closed-book models such as Flan-PaLM (540B) and Med-PaLM 2. This showcases the potential of combining large language models with targeted domain expertise, emphasizing the value of domain-specific knowledge in retrieval-augmented generation methods.

3.4.1 Component Impact Analysis

Our investigation into the LLM-AMT pipeline reveals the integral roles of the Textbook Retriever, Query Augmenter, and Knowledge Self-Refiner. In Table 4, we provide a unified analysis, demonstrating their collective impact on enhancing the model's performance on medical QA tasks, as evidenced in the MedQA-USMLE dataset and corroborated by similar trends in other datasets.

1. **Textbook Retriever (HybTextR)** serves as the cornerstone, providing a 7.3% boost in accuracy by tapping into specialized medical literature for relevant information.

2. **Query Augmenter** elevates recall by translating general inquiries into precise medical terminology and through query expansion to enhance relevant knowledge association, leading to a 3.4% incremental accuracy gain. It ensures that the breadth of the search captures a wide spectrum of relevant evidence.

3. **Knowledge Self-Refiner** complements by scrutinizing the relevance and usefulness of the retrieved information, fine-tuning precision, and contributing a further 1.9% accuracy increase. It filters the evidence, sharpening the focus on the most pertinent medical facts.

4. **Synergistic Effect**: The Query Augmenter and Knowledge Self-Refiner synergize to elevate LLM performance. The augmenter boosts knowledge recall, while the refiner improves precision, providing the LLM with high-quality, external medical knowledge. This synergy is crucial for handling complex

Method	Accuracy
GPT-3.5-Turbo	51.3
+ retriever	58.6
+ query rewriting	61.2
+ query expansion	62.0

Table 5: Query Augmenter's ablation study

	Zero-shot	Fine-tuned
	MedQA-USMLE	MedQA-USMLE
BM25	55.6	–
Sparse	57.4	59.3
Dense	59.7	60.9
ColBERT	58.2	61.5
Sparse + Dense	60.1	62.7
Sparse + Rerank	59.5	61.3
Dense + Rerank	60.6	63.7
HybTextR	**62.0**	**64.1**

Table 6: Evaluation of Retrieval and Reranking Strategies on the Performance of LLM-AMT

medical queries. See the appendix A.7 for a detailed case study on their interaction.

3.5 Ablation Study

Here, we perform ablation studies on the query augmentation, the retrieval mechanisms, and the knowledge self-refinement strategy to refine and identify the most optimal configuration specifically tailored for question-answering tasks within the medical domain.

3.5.1 Query Augmenter Components

In this part, we performed a series of ablation experiments on the MedQA dataset to evaluate the efficacy of various components within our Query Augmenter framework. The detailed results are presented in Table 5. Our findings demonstrate that each component contributes to the overall performance incrementally. These results underscore the synergistic effect of these components in improving the model's ability to understand and process complex medical queries.

3.5.2 Textbook Retrievers

In Table 6, we evaluate the impact of different retrieval methods in our pipeline. The late-interaction ColBERT retriever notably achieves 58.2% accuracy on MedQA-USMLE in a zero-shot scenario, surpassing standalone sparse and dense retrievers. A hybrid approach, combining dense and sparse retrievers, yields a higher accuracy of 60.1%.

Adding a reranker, the Dense + Rerank setup increases accuracy to 60.6%. The HybTextR model, incorporating sparse, dense, and reranking, reaches the peak accuracy of 62.0% on MedQA-USMLE, demonstrating the advantage of a layered retrieval approach in medical contexts. For similar experiments on MedQA-MCMLE and MedMCQA, refer to Appendix A.1.

3.5.3 Knowledge Self-Refiner Components

Configuration	Accuracy (%)
w/o KSR	64.1
+ Relevance Filter	65.8
+ Usefulness Filter	66.2
+ Full System	67.9

Table 7: Impact of KSR components on GPT-3.5-Turbo LLM Reader accuracy.

In this part, the impact of Knowledge Self-Refiner (KSR) components on a GPT-3.5-Turbo LLM Reader is examined within the MedQA domain, supplemented by HybTextR and Query Augmentation strategies. As illustrated in Table 7, the Relevance Filter marginally increases accuracy, underscoring its role in identifying pertinent content. The Usefulness Filter, contributing a slightly higher gain, is instrumental in isolating content of practical value for responses. The concurrent application of both filters results in the highest accuracy, signifying the importance of multi-dimensional content refinement in medical question-answering.

3.6 Further Discussion

In this section, we discuss and further assess our models, particularly their performance in non-multiple-choice medical QA tasks.

Tiers	GPT-3.5	LLM-AMT
Correct	27	36
Mostly Correct	10	12
Partially Correct	14	19
Wrong	49	33

Table 8: Evaluation of the Non-multiple-choice Medical Question Answering Task. GPT-3.5 as the LLM Reader.

To test medical QA models in a realistic scenario, we chose 100 varied questions from the MedQA-USMLE dataset and produced answers without seeing the options. Medical professionals evaluated the answer quality, ranking them as:

- **Correct**: Accurate and complete.

- **Mostly Correct**: Generally accurate, with some details missing.
- **Partially Correct**: Contains correct aspects but lacks key information.
- **Wrong**: Inaccurate or irrelevant.

Our LLM-AMT model surpassed the GPT-3.5-Turbo baseline in the non-multiple-choice QA task, delivering 36 correct answers to the baseline's 27. Notably, LLM-AMT provided more partially correct answers (19 vs. 14) and fewer errors (33 vs. 49). This underscores the model's enhanced accuracy in the medical QA domain, as detailed in Table 8. The superior performance of LLM-AMT in the non-multiple-choice QA task not only illustrates its advanced capabilities but also emphasizes its potential for practical application in real-world medical scenarios. Such advancements can be instrumental in aiding medical professionals with more accurate and reliable information.

4 Related Work

In this section, we provide an overview of the related work in biomedical QA, retrieval-augmented QA, and text retrieval.

4.1 Biomedical question answering

Biomedical QA plays a pivotal role in clinical decision support (Ely et al., 2005) and the acquisition of biomedical knowledge (Jin et al., 2022). With the rise of pre-trained language models (LMs), there's been a significant uptick in performance and the emergence of new capabilities across various natural language processing (NLP) tasks (Chowdhery et al., 2022; Chung et al., 2022; Wei et al., 2022b,a). Nevertheless, these auto-regressive LLMs, when applied in domains like medicine and healthcare that require intensive knowledge or reasoning, are prone to generating hallucinations and erroneous content. Combining external knowledge sources with LLMs is a promising approach to counteract these pitfalls (Mialon et al., 2023).

4.2 Retrieval Augmented Generation

The retrieval-augmented generation paradigm, originating from the DrQA framework by Chen et al., initially used heuristic retrievers like TF-IDF to source evidence from Wikipedia, followed by a neural model to extract answers. This methodology was advanced by DPR (Karpukhin et al., 2020), using pre-trained transformers like BERT for retrieval and reading. Retrieval Augmented Generation (RAG) (Lewis et al., 2020) further evolved the approach by shifting from answer extraction to generation, enabling free-form text creation. Advances in RAG have explored retrieval as a critical tool for augmentation, with Schick et al., Luo et al., and Asai et al. targeting enhanced information sourcing mechanisms. Moreover, Yan et al. explored and designed corrective strategies for RAG to bolster generation robustness. Concurrently, models like REALM (Guu et al., 2020) and RETRO (Borgeaud et al., 2022) integrated retrieval during the pre-training phase. Recently, Large Language Models (LLMs) have been incorporated into this framework, as seen in REPLUG (Shi et al., 2023b) and IC-RALM (Ram et al., 2023). While prior work on RAG primarily addressed general knowledge, this study introduces the first application of RAG to medical literature, harnessing a vast collection of medical textbooks. Our innovative knowledge self-refinement strategies enhance the fidelity of retrieved information, marking the first refinement of RAG's retrieval component for elevated performance in the medical domain.

4.3 Neural Text Retrieval

Recent progress in Neural Retrieval with bi-encoder architectures surpasses traditional methods like BM25/TF-IDF. This technique encodes queries and documents independently using pre-trained transformers, measuring similarity with embedding distances. Neural retrieval can be categorized into *dense retrieval* (e.g., DPR (Karpukhin et al., 2020), ANCE (Xiong et al., 2020), CoCondenser (Gao and Callan, 2021)), *sparse retrieval* (e.g., DeepImpact (Mallia et al., 2021), uniCOIL (Lin and Ma, 2021), SPLADE (Formal et al., 2021)), and *late interaction retrieval* (e.g., ColBERT (Khattab and Zaharia, 2020), COIL (Gao et al., 2021)), based on the type of embedding used. In this study, we apply these neural retrieval methods to medical textbook retrieval, assessing their domain-specific effectiveness beyond standard corpora.

5 Conclusion

We introduced LLM-AMT, a novel pipeline optimized for medical tasks, harnessing authoritative medical textbooks to enhance LLMs' accuracy and professionalism. Empirical evaluations reinforced the value of integrating domain-specific textbooks

with LLMs, providing an avenue for future studies. Further, our ablation study delineated the significance of external knowledge retrieval, query augmentation, and knowledge self-refinement strategy within our proposed architecture. These findings set a precedent for advancing specialized domain-aware models, especially in the context of medical informatics and healthcare AI applications.

6 Limitations

6.1 Model Explainability

One significant limitation of the LLM-AMT system lies in its inherent lack of explainability. While the integration of authoritative medical textbooks enhances the model's responses, the reasoning behind these responses often remains opaque. Medical decision-making demands a high degree of transparency; however, as with many large language models, the LLM-AMT operates as a "black box". This poses a challenge in clinical settings, where explanations for diagnoses or treatment recommendations are crucial for trust and accountability. Incorrect or unexplained advice from the model could lead to misdiagnosis or inappropriate treatment, endangering patient health and potentially eroding trust in AI-assisted medical systems. The model's inability to provide detailed explanations for its conclusions can be a significant barrier to its adoption in practice.

6.2 Interactive Question-Answering

Another limitation is the system's capacity for interactive QA. In real-world medical practice, diagnostic and treatment processes involve nuanced communications with patients, requiring a deep understanding of individual circumstances, empathetic engagement, and the ability to ask follow-up questions for clarification. The LLM-AMT, despite its advancements, cannot fully replicate this level of interaction. The model might not adequately handle the subtleties of patient-specific narratives or the dynamic nature of medical conversations. Thus, while LLM-AMT can provide informative responses, its interactive capabilities are limited in comparison to the rich, two-way communication typically found in patient-clinician interactions.

References

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

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

John W Ely, Jerome A Osheroff, M Lee Chambliss, Mark H Ebell, and Marcy E Rosenbaum. 2005. Answering physicians' clinical questions: obstacles and potential solutions. *Journal of the American Medical Informatics Association*, 12(2):217–224.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. Splade: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2288–2292.

Jianfeng Gao, Chenyan Xiong, Paul Bennett, and Nick Craswell. 2022a. Neural approaches to conversational information retrieval. *arXiv preprint arXiv:2201.05176*.

Luyu Gao and Jamie Callan. 2021. Condenser: a pre-training architecture for dense retrieval. *arXiv preprint arXiv:2104.08253*.

Luyu Gao, Zhuyun Dai, and Jamie Callan. 2021. Coil: Revisit exact lexical match in information retrieval with contextualized inverted list. *arXiv preprint arXiv:2104.07186*.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2022b. Precise zero-shot dense retrieval without relevance labels. *arXiv preprint arXiv:2212.10496*.

Luyu Gao, Xueguang Ma, Jimmy J. Lin, and Jamie Callan. 2022c. Tevatron: An efficient and flexible toolkit for dense retrieval. *ArXiv*, abs/2203.05765.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

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, Zheng Yuan, Guangzhi Xiong, Qianlan Yu, Huaiyuan Ying, Chuanqi Tan, Mosha Chen, Songfang Huang, Xiaozhong Liu, and Sheng Yu. 2022. Biomedical question answering: a survey of approaches and challenges. *ACM Computing Surveys (CSUR)*, 55(2):1–36.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

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

Valentin Liévin, Christoffer Egeberg Hother, and Ole Winther. 2022. Can large language models reason about medical questions? *arXiv preprint arXiv:2207.08143*.

Valentin Liévin, Andreas Geert Motzfeldt, Ida Riis Jensen, and Ole Winther. 2023. Variational open-domain question answering. In *International Conference on Machine Learning*, pages 20950–20977. PMLR.

Jimmy Lin and Xueguang Ma. 2021. A few brief notes on deepimpact, coil, and a conceptual framework for information retrieval techniques. *arXiv preprint arXiv:2106.14807*.

Hongyin Luo, Yung-Sung Chuang, Yuan Gong, Tianhua Zhang, Yoon Kim, Xixin Wu, Danny Fox, Helen Meng, and James Glass. 2023. Sail: Search-augmented instruction learning. *arXiv preprint arXiv:2305.15225*.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409.

Antonio Mallia, Omar Khattab, Torsten Suel, and Nicola Tonellotto. 2021. Learning passage impacts for inverted indexes. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1723–1727.

Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, et al. 2023. Augmented language models: a survey. *arXiv preprint arXiv:2302.07842*.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. *choice*, 2640:660.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale

multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning*, pages 248–260. PMLR.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *arXiv preprint arXiv:2302.00083*.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023a. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023b. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2022. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Chao-Hong Tan, Jia-Chen Gu, Chongyang Tao, Zhen-Hua Ling, Can Xu, Huang Hu, Xiubo Geng, and Daxin Jiang. 2022. Tegtok: Augmenting text generation via task-specific and open-world knowledge. *arXiv preprint arXiv:2203.08517*.

Liang Wang, Nan Yang, and Furu Wei. 2023. Query2doc: Query expansion with large language models. *arXiv preprint arXiv:2303.07678*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Bohong Wu, Zhuosheng Zhang, Jinyuan Wang, and Hai Zhao. 2021. Sentence-aware contrastive learning for open-domain passage retrieval. *arXiv preprint arXiv:2110.07524*.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. Linkbert: Pretraining language models with document links. *arXiv preprint arXiv:2203.15827*.

Xiao Zhang, Ji Wu, Zhiyang He, Xien Liu, and Ying Su. 2018. Medical exam question answering with large-scale reading comprehension. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

A Appendix

A.1 Exclusive Performance Evaluation on MedQA-MCMLE and MedMCQA

Table 10 presents the performance of various retrieval and reranking strategies exclusively on the MedQA-MCMLE and MedMCQA datasets. These datasets pose distinct challenges compared to MedQA-USMLE and thus merit a separate analysis. The results provide insights into the generalizability and robustness of the methods when confronted with different types of medical question-answering datasets. The HybTextR method, in particular, shows a consistently strong performance, suggesting its potential as a versatile tool for medical information retrieval tasks.

A.2 Fine-tuning Hyperparameters for Retrievers

Table 9 presents the hyperparameters used for fine-tuning retrievers. In the *Seq Length* column of Table 9, the notation $32 + 220$ for the ColBERT model indicates that the maximum length for the query is set to 32, while the length for the passage is 220. For the Reranker model, the input sequence is structured as `[CLS]` token followed by the query, then a `[SEP]` token, and finally the

passage. Therefore, its sequence length is calculated as $1 + 126 + 1 + 384$, which sums up to 512.

Model	Batch	Seq Len	LR
Splade	64	256	2×10^{-5}
DPR	8	256	1×10^{-5}
ColBERT	32	32+220	3×10^{-6}
Reranker	8	126+384	8×10^{-6}

Table 9: Hyperparameters for fine-tuning.

A.3 Models

We use the following model:

- **DPR**, which uses BERT-base as the backbone and has 110M parameters. It is under the CC-BY-NC 4.0 License.
- **SPLADE**, which uses BERT-base as the backbone and has 110M parameters. It is under the CC BY-NC-SA 4.0. License.
- **ColBERT**, which uses BERT-base as the backbone and has 110M parameters. It is under the MIT License.
- **LLaMA-2-13B**, 13B parameters, under the Llama 2 Community License Agreement.
- **GPT-3.5-Turbo and GPT-4-Turbo**, which are not open-source and can only be accessed via API requests.

A.4 Datasets

We use the following datasets:

- **MedQA**, which is under the MIT Licenses. The intended purpose of the MedQA dataset is to support and advance research in the area of natural language processing (NLP) and information retrieval (IR) within the medical domain. MedQA is composed of both English and Chinese questions and answers. While the dataset predominantly features clinical scenarios and medical knowledge representations, demographic information of the represented groups is not explicitly detailed due to the nature of the data.
- **MedMCQA**, which is under the MIT License for non-commercial research purposes. The MedMCQA dataset spans a broad range of medical domains, including but not limited to cardiology, oncology, pediatrics, neurology, and infectious diseases. Each domain is represented with questions and answers that reflect the diversity of medical knowledge. The dataset is primarily in English, ensuring that the findings of our research are directly applicable to English-language medical question-answering systems.

A.5 AI Assistance in Writing

In the preparation of this manuscript, we utilized an AI language model, specifically ChatGPT, to assist with grammar checking and refining the expressions used in our writing. This utilization was confined to ensuring linguistic accuracy and enhancing readability, without influencing the scientific content or the originality of the research findings presented. The contribution of ChatGPT was strictly as a supportive tool for language polishing, and all final decisions regarding the manuscript content were made by the human authors.

A.6 Full List of Instructions For GPT-3.5-Turbo

In Table 11, we list the instructions we used in LLM-AMT.

A.7 Case Study

As part of our comprehensive case study, Tables 12, 13, 14, and 15 present detailed input and output data corresponding to each component depicted in the overview (Figure 1). These tables include the full question and options, retrieved passages from the textbook, retrieved passages from Wikipedia, the rewritten query, the expanded query, the results from the knowledge self-refinement stage, and the final refined knowledge. This granular view provides clear insight into the information processing pipeline and the effectiveness of each module.

	Zero-shot		Fine-tuned	
	MedQA-MCMLE	**MedMCQA**	**MedQA-MCMLE**	**MedMCQA**
BM25	59.7	55.2	–	–
Sparse	60.4	57.5	62.9	59.6
Dense	61.0	57.7	63.8	59.3
ColBERT	62.4	58.1	64.1	60.4
Sparse + Dense	64.9	58.7	65.5	61.9
Sparse + Rerank	63.8	59.2	65.2	62.8
Dense + Rerank	**65.4**	61.8	65.3	64.6
HybTextR	64.4	**63.1**	**68.9**	**65.2**

Table 10: Evaluation of Retrieval and Reranking Strategies on MedQA-MCMLE and MedMCQA Datasets

Query Rewriting: Question: XXX Please reformulate the given question by employing precise medical terminology. Focus on capturing the essence of the patient's symptoms and conditions in a generalized form that reflects common clinical descriptions. Avoid using colloquial language and ensure that the rewritten query is clear, concise, and can be universally understood in a professional medical context.
Query Expansion: Question: XXX Assume the role of a medical doctor and expand upon the initial query. Conduct a systematic analysis by dissecting the question into its medical components. Then, elaborate on each component with detailed medical insights that collectively build a comprehensive understanding of the underlying health issue. Proceed methodically to ensure that each step of your explanation contributes to a logically structured answer.
Relevance Filter: Retrieved Passage: XXX x: XXX Examine the retrieved passages above carefully. Determine if each passage pertains to the context of the specific query represented by 'x'. Respond with 'Yes' if a passage is relevant and contributes meaningful information to the query, or 'No' if it does not relate to the query or provide valuable insight. Please answer with 'Yes' or 'No' only for each passage assessed.
Usefulness Filter: Retrieved Information: XXX x: XXX Review the information retrieved above and evaluate its utility in addressing the question represented by 'x'. Provide a response of 'Yes' if the information is pertinent and aids in formulating a comprehensive answer, or 'No' if it lacks relevance or does not contribute to a substantive response to the question. Respond with a singular 'Yes' or 'No' for the usefulness of each piece of information.
LLM Reader Instruction: Medical Knowledge: XXX Question: XXX Using the medical knowledge provided, please answer the following medical question with a chain-of-thought approach. Break down your reasoning into clear, logical steps that detail your clinical thought process from initial hypothesis formation through to the final conclusion, similar to how a medical professional would approach a diagnostic challenge. Your answer should not only be informed by the medical knowledge but also transparent in the reasoning that led to your conclusion.

Table 11: Full List of Instructions For GPT-3.5-Turbo

Question:
Two weeks after undergoing an emergency cardiac catheterization with stenting for unstable angina pectoris, a 61-year-old man presents with decreased urinary output and malaise. He has type 2 diabetes mellitus and osteoarthritis of the hips. His medications prior to admission were insulin and naproxen, and he was started on aspirin, clopidogrel, and metoprolol after the coronary intervention. His current vitals are: temperature 38°C (100.4°F), pulse 93/min, blood pressure 125/85 mm Hg. Examination reveals mottled, reticulated purplish discoloration of the feet. Laboratory studies show: Hemoglobin count 14 g/dL, Leukocyte count 16,400/mm3, Segmented neutrophils 56%, Eosinophils 11%, Lymphocytes 31%, Monocytes 2%, Platelet count 260,000/mm3, Erythrocyte sedimentation rate 68 mm/h, Serum Urea nitrogen 25 mg/dL, Creatinine 4.2 mg/dL. Renal biopsy shows intravascular spindle-shaped vacuoles. The question is: Which of the following is the most likely cause of this patient's symptoms?
Options:
A) Renal papillary necrosis
B) Cholesterol embolization
C) Eosinophilic granulomatosis with polyangiitis
D) Polyarteritis nodosa

Ground Truth:
Cholesterol embolization (Choice B)

Rewritten Query:
What is the most likely etiology of the oliguria and malaise in this patient with a history of emergency cardiac catheterization, type 2 diabetes mellitus, and **osteoarthritis of the hips**, along with **mottled**, **reticulated purplish discoloration of the feet**, **elevated creatinine levels**, and **intravascular spindle-shaped vacuoles on renal biopsy**?

Expanded Query:
Given the clinical scenario and lab findings, we should consider several factors to systematically approach this medical question. Let's analyze each option step by step.
A. Renal Papillary Necrosis: Presence of diabetes could predispose the patient to this condition, but the biopsy's intravascular spindle-shaped vacuoles are not characteristic of renal papillary necrosis, and the discoloration of the feet isn't a typical symptom.
B. Cholesterol Embolization: Recent cardiac catheterization could lead to this condition; however, one would need to consider other signs and symptoms and correlate them with further tests to rule out or confirm this diagnosis. The biopsy finding and the patient's presentation could be suggestive, but are not definitive without further evidence.
C. Eosinophilic Granulomatosis with Polyangiitis (EGPA): While the patient has eosinophilia, the absence of asthma or typical respiratory symptoms and granuloma formation on biopsy makes this option less likely.
D. Polyarteritis Nodosa (PAN): The absence of typical symptoms and biopsy findings associated with PAN, alongside the presence of eosinophilia, which isn't commonly seen in PAN, makes this diagnosis less likely.
In summary, while certain features may seem to align with some of the conditions, the evidence is not conclusive for any of the options based on the information presented. Further diagnostic workup would be essential to come to a definitive diagnosis.

Table 12: Case Study

Retrieved Passages from Wikipedia:

Cardiac catheterization is a medical procedure used to diagnose and treat certain cardiovascular conditions. During cardiac catheterization, a long thin tube called a catheter is inserted in an artery or vein in your groin, neck, or arm and threaded through your blood vessels to your heart. Using this catheter, doctors can then perform diagnostic tests as part of a cardiac cath work-up. Procedures can include coronary angiography, as well as treatments like coronary angioplasty or stenting to open up narrowed or blocked arteries. ✗

Type 2 diabetes mellitus is a chronic metabolic disorder characterized by high blood sugar, insulin resistance, and relative lack of insulin. It is often associated with obesity and a sedentary lifestyle and can lead to severe health complications if not managed effectively. Management includes lifestyle changes, monitoring blood sugar levels, and medication or insulin therapy. ✗

Osteoarthritis of the hips is a degenerative joint disease that affects the hip joints. It occurs when the protective cartilage that cushions the tops of the bones wears down over time, which can lead to pain, swelling, and reduced range of motion in the hip. While the exact cause is unknown, factors such as age, joint injury, obesity, and genetics can increase the risk of developing hip osteoarthritis. ✗

Mottled, reticulated purplish discoloration of the feet is a skin condition that can indicate a variety of circulatory problems. The appearance of a lace-like pattern of purplish discoloration, known as livedo reticularis, may be due to the occlusion of small blood vessels or could be a reaction to cold temperatures. In certain cases, it can be associated with more serious underlying conditions, such as cholesterol embolization syndrome or vasculitis. ✗

Elevated creatinine levels in the blood can indicate impaired kidney function or kidney disease. Creatinine is a waste product produced by muscles from the breakdown of a compound called creatine. Normally, the kidneys filter out creatinine from the blood and excrete it in urine. High levels of creatinine may signal that the kidneys are not working properly. ✗

Intravascular spindle-shaped vacuoles observed on renal biopsy can be indicative of a pathological process affecting the kidneys. These structures could suggest the presence of intravascular material, such as cholesterol emboli, which can result from the dislodgement of atheromatous plaques following procedures like cardiac catheterization. The presence of such vacuoles requires further pathological examination to determine the precise cause and appropriate treatment. ✗

...

Retrieved Passages from Textbooks:

Cholesterol embolization, also known as cholesterol crystal embolism (CCE) or atheroembolism, is a pathology that arises from the showering of cholesterol crystals from atherosclerotic plaques into the systemic circulation. The condition can result in a vasculopathy that is often systemic and can lead to end-organ damage due to the occlusion of small to medium-sized vessels. Pathophysiology. Cholesterol embolization occurs when cholesterol crystals dislodge from plaques in larger arteries, such as the aorta, and travel through the bloodstream to smaller blood vessels, where they become lodged and cause ischemia. The release of cholesterol crystals can occur spontaneously, but it is most commonly associated with iatrogenic causes such as cardiac catheterization, vascular surgery, and anticoagulation therapy. Clinical Presentation. The clinical manifestations of cholesterol embolization are variable and depend on the organ systems involved. Common symptoms include: Cutaneous: Livedo reticularis, cyanotic toes, or gangrene due to skin and soft tissue ischemia. Renal: Acute or chronic renal insufficiency presenting as an increase in serum creatinine, proteinuria, and hematuria. Gastrointestinal: Abdominal pain, gastrointestinal bleeding, and pancreatitis. Neurological: Transient ischemic attacks, strokes, and multi-infarct dementia. ✓

Table 13: Continued Case Study

Retrieved Passages from Textbook:

Diagnosis of Cholesterol embolization is often challenging due to the nonspecific nature of the symptoms. Cholesterol embolization is frequently underdiagnosed or misdiagnosed. Laboratory findings may include eosinophilia and elevated erythrocyte sedimentation rate (ESR). Definitive diagnosis typically requires histological examination of affected tissue, which would reveal biconvex, needle-shaped clefts within the vessel lumen, indicative of cholesterol crystals. There is no specific treatment for cholesterol embolization. Management is primarily supportive and focuses on treating complications and symptomatic relief. Statin therapy may be used to stabilize atherosclerotic plaques and reduce the risk of further embolization. In cases related to medication, such as anticoagulants or thrombolytics, discontinuation or adjustment of the drugs may be required. The prognosis for patients with cholesterol embolization varies and can range from benign to severe, life-threatening multiorgan failure. Early recognition and management of the condition are crucial to improving outcomes. ✓

Cardiac Catheterization Cardiac catheterization is an invasive diagnostic procedure that provides comprehensive information about the heart and surrounding blood vessels. This procedure involves the insertion of a catheter, which is a thin, flexible tube, into the coronary arteries and heart chambers under X-ray guidance. Indications Cardiac catheterization is typically indicated for the following purposes: Diagnosing coronary artery disease Evaluating heart muscle function Determining the need for further treatment (such as angioplasty or coronary artery bypass surgery) Assessing the effectiveness of prior procedures Measuring the pressure within the heart chambers Diagnosing congenital heart abnormalities ✗

...

Passages after Relevance Filter:

Cholesterol embolization, also known as cholesterol crystal embolism (CCE) or atheroembolism, is a pathology that arises from the showering of cholesterol crystals from atherosclerotic plaques into the systemic circulation. The condition can result in a vasculopathy that is often systemic and can lead to end-organ damage due to the occlusion of small to medium-sized vessels. Pathophysiology. Cholesterol embolization occurs when cholesterol crystals dislodge from plaques in larger arteries, such as the aorta, and travel through the bloodstream to smaller blood vessels, where they become lodged and cause ischemia. The release of cholesterol crystals can occur spontaneously, but it is most commonly associated with iatrogenic causes such as cardiac catheterization, vascular surgery, and anticoagulation therapy. Clinical Presentation. The clinical manifestations of cholesterol embolization are variable and depend on the organ systems involved. Common symptoms include: Cutaneous: Livedo reticularis, cyanotic toes, or gangrene due to skin and soft tissue ischemia. Renal: Acute or chronic renal insufficiency presenting as an increase in serum creatinine, proteinuria, and hematuria. Gastrointestinal: Abdominal pain, gastrointestinal bleeding, and pancreatitis. Neurological: Transient ischemic attacks, strokes, and multi-infarct dementia.

Diagnosis of Cholesterol embolization is often challenging due to the nonspecific nature of the symptoms. Cholesterol embolization is frequently underdiagnosed or misdiagnosed. Laboratory findings may include eosinophilia and elevated erythrocyte sedimentation rate (ESR). Definitive diagnosis typically requires histological examination of affected tissue, which would reveal biconvex, needle-shaped clefts within the vessel lumen, indicative of cholesterol crystals. There is no specific treatment for cholesterol embolization. Management is primarily supportive and focuses on treating complications and symptomatic relief. Statin therapy may be used to stabilize atherosclerotic plaques and reduce the risk of further embolization. In cases related to medication, such as anticoagulants or thrombolytics, discontinuation or adjustment of the drugs may be required. The prognosis for patients with cholesterol embolization varies and can range from benign to severe, life-threatening multiorgan failure. Early recognition and management of the condition are crucial to improving outcomes.

Table 14: Continued Case Study

Segments after Usefulness Filter:

Cholesterol embolization, also known as cholesterol crystal embolism (CCE) or atheroembolism, is a pathology that arises from the showering of cholesterol crystals from atherosclerotic plaques into the systemic circulation. The condition can result in a vasculopathy that is often systemic and can lead to end-organ damage due to the occlusion of small to medium-sized vessels. ✓

Pathophysiology. Cholesterol embolization occurs when cholesterol crystals dislodge from plaques in larger arteries, such as the aorta, and travel through the bloodstream to smaller blood vessels, where they become lodged and cause ischemia. The release of cholesterol crystals can occur spontaneously, but it is most commonly associated with iatrogenic causes such as cardiac catheterization, vascular surgery, and anticoagulation therapy. ✓

Clinical Presentation. The clinical manifestations of cholesterol embolization are variable and depend on the organ systems involved. Common symptoms include: Cutaneous: Livedo reticularis, cyanotic toes, or gangrene due to skin and soft tissue ischemia. Renal: Acute or chronic renal insufficiency presenting as an increase in serum creatinine, proteinuria, and hematuria. Gastrointestinal: Abdominal pain, gastrointestinal bleeding, and pancreatitis. Neurological: Transient ischemic attacks, strokes, and multi-infarct dementia. ✓

Diagnosis of Cholesterol embolization is often challenging due to the nonspecific nature of the symptoms. Cholesterol embolization is frequently underdiagnosed or misdiagnosed. Laboratory findings may include eosinophilia and elevated erythrocyte sedimentation rate (ESR). Definitive diagnosis typically requires histological examination of affected tissue, which would reveal biconvex, needle-shaped clefts within the vessel lumen, indicative of cholesterol crystals. ✓

There is no specific treatment for cholesterol embolization. Management is primarily supportive and focuses on treating complications and symptomatic relief. Statin therapy may be used to stabilize atherosclerotic plaques and reduce the risk of further embolization. In cases related to medication, such as anticoagulants or thrombolytics, discontinuation or adjustment of the drugs may be required. ✗

The prognosis for patients with cholesterol embolization varies and can range from benign to severe, life-threatening multiorgan failure. Early recognition and management of the condition are crucial to improving outcomes. ✗

Refined Knowledge:

Cholesterol embolization, also known as cholesterol crystal embolism (CCE) or atheroembolism, is a pathology that arises from the showering of cholesterol crystals from atherosclerotic plaques into the systemic circulation. The condition can result in a vasculopathy that is often systemic and can lead to end-organ damage due to the occlusion of small to medium-sized vessels. Pathophysiology. Cholesterol embolization occurs when cholesterol crystals dislodge from plaques in larger arteries, such as the aorta, and travel through the bloodstream to smaller blood vessels, where they become lodged and cause ischemia. The release of cholesterol crystals can occur spontaneously, but it is most commonly associated with iatrogenic causes such as cardiac catheterization, vascular surgery, and anticoagulation therapy. Clinical Presentation. The clinical manifestations of cholesterol embolization are variable and depend on the organ systems involved. Common symptoms include: Cutaneous: Livedo reticularis, cyanotic toes, or gangrene due to skin and soft tissue ischemia. Renal: Acute or chronic renal insufficiency presenting as an increase in serum creatinine, proteinuria, and hematuria. Gastrointestinal: Abdominal pain, gastrointestinal bleeding, and pancreatitis. Neurological: Transient ischemic attacks, strokes, and multi-infarct dementia. Diagnosis of Cholesterol embolization is often challenging due to the nonspecific nature of the symptoms. Cholesterol embolization is frequently underdiagnosed or misdiagnosed. Laboratory findings may include eosinophilia and elevated erythrocyte sedimentation rate (ESR). Definitive diagnosis typically requires histological examination of affected tissue, which would reveal biconvex, needle-shaped clefts within the vessel lumen, indicative of cholesterol crystals.

Table 15: Continued Case Study