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

Anonymous submission

Abstract

Large-scale language models (LLMs), such as ChatGPT, are capable of generating human-like responses for various downstream tasks, such as task-oriented dialogues and question answering. However, applying LLMs to medical domains remains challenging due to their inability to leverage domain-specific knowledge. In this study, we present the **Large-scale Language Models Augmented with Medical Textbooks (LLM-AMT)**, which integrates authoritative medical textbooks as the cornerstone of its design, enhancing its proficiency in the specialized domain through plug-and-play modules, comprised of a Hybrid Textbook Retriever, supplemented by the Query Augmenter and the LLM Reader. Experimental evaluation on three open-domain medical question-answering tasks reveals a substantial enhancement in both the professionalism and accuracy of the LLM responses when utilizing LLM-AMT, exhibiting an improvement ranging from 11.4% to 13.2%. Despite being 100× smaller, we found that medical textbooks as the retrieval corpus serves as a more valuable external knowledge source than Wikipedia in the medical domain. Our experiments show that textbook augmentation results in a performance improvement ranging from 9.7% to 12.2% over Wikipedia augmentation. We will release our code based on acceptance.

1 Introduction

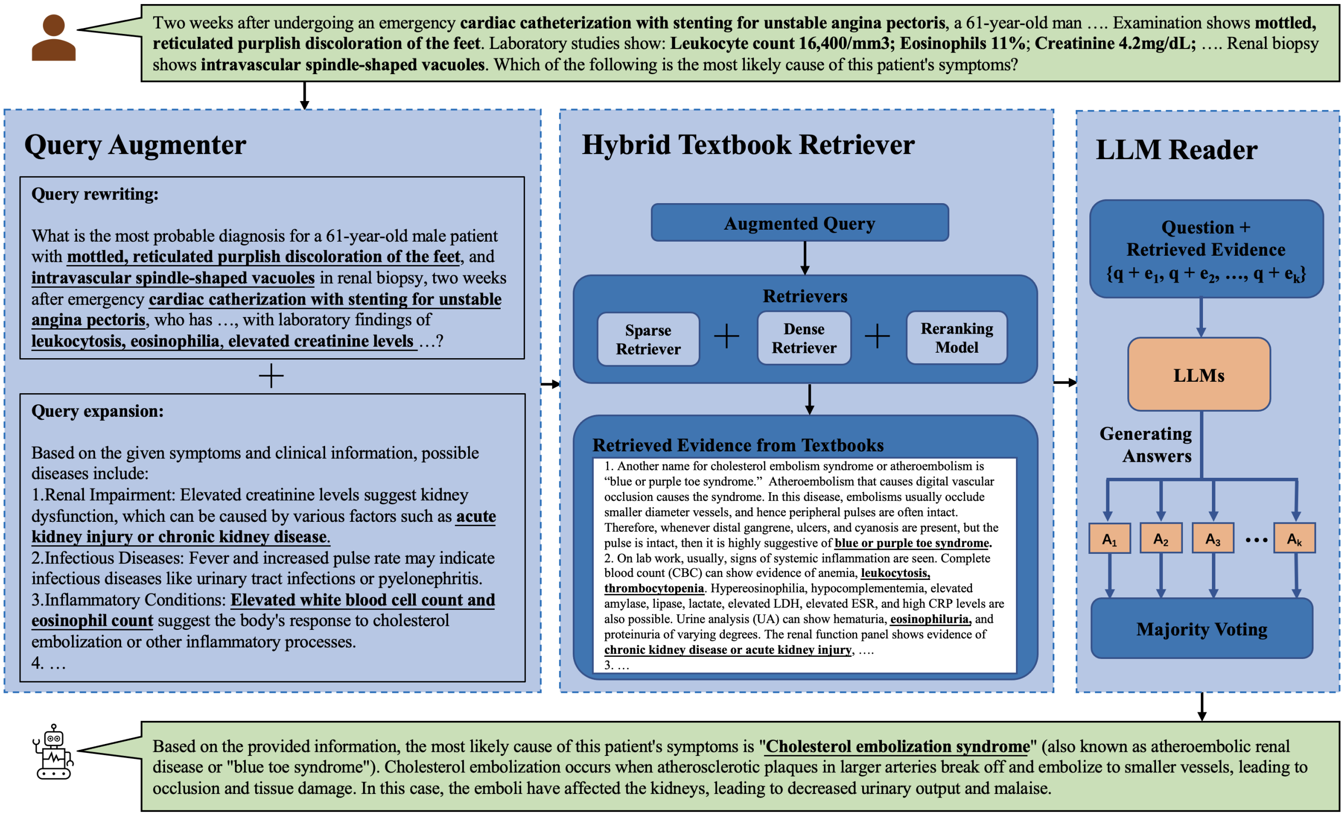
Language forms the foundation of health and medicine, facilitating interactions between individuals and healthcare providers. Recent advancements in Large Language Models (LLMs) have opened up new possibilities for exploring the potential of artificial intelligence (AI) systems in the medical domain, enabling them to comprehend and communicate through language. This progress holds great promise for enhancing human-AI collaboration, as evidenced by the impressive performance of these models on various medical question answering datasets (Zhang et al. 2018; Pal, Umapathi, and Sankarasubbu 2022; Jin et al. 2019).

The existing LLMs are mainly trained to encode all the world knowledge implicitly into the parameter space. However, the knowledge encoding process in LLMs may lead to information loss and "memory distortion" (Peng et al. 2023), causing these models to potentially generate plausible-sounding but incorrect content (hallucinate). This characteristic can pose significant challenges, particularly when such models are applied to critical tasks. Recently, there is increasing interest in augmenting LLMs with external knowledge to deal with this issue, but most of the previously proposed approaches necessitate fine-tuning the parameters of LLMs (e.g., Luo et al.; Gao et al.; Singhal et al.), which can become prohibitively costly as the LLM size grows exponentially.

The *retrieve-then-read* architecture (Lewis et al. 2020; Karpukhin et al. 2020; Izacard et al. 2022) has emerged as an efficient alternative to compute-intensive fine-tuning. In open-domain QA, a retriever identifies relevant documents for an input question, and then a reader extracts the answer using these documents as context. Prior works either enhanced the retrieval process (Wu et al. 2021; Izacard et al. 2021) or jointly fine-tuned the reader model (Lewis et al. 2020; Izacard et al. 2022). More recent advances utilize a powerful LLM as the reader, emphasizing LLM-oriented adaptation (Shi et al. 2023). 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 in complex domains.

In this paper, we propose LLM-AMT, an innovative approach tailored to the medical domain. Our primary aim is to leverage *textbooks* as a rich external knowledge source, thereby enhancing the capabilities of Large Language Models (LLMs) in the medical field. As illustrated in Figure 1, given an input medical question, the Query Augmenter utilizes LLMs to rewrite and expand the original question, generating more informative queries. Provided with the augmented query, the Textbook Retriever can integrate different types of retrievers to recall relevant evidence from the plain text of medical textbooks. Finally, the LLM Reader takes the question and retrieved evidence as context to obtain the responses. On downstream QA tasks, we can obtain multiple responses and then combine them with the majority voting to derive the final answer.

To demonstrate the effectiveness of the proposed method in the medical domain, we conduct empirical evaluations of LLM-AMT on three datasets: MedQA-USMLE, MedQA-MCMLE, and MedMCQA. We use GPT-3.5 as the baseline for our comparisons. Our results reveal that, using LLM-AMT, LLM responses witnessed substantial enhancements over the GPT-3.5 baseline, with improvements spanning



Two weeks after undergoing an emergency **cardiac catheterization with stenting for unstable angina pectoris**, a 61-year-old man …. Examination shows **mottled, reticulated purplish discoloration of the feet**. Laboratory studies show: **Leukocyte count 16,400/mm3; Eosinophils 11%**; **Creatinine 4.2mg/dL;** …. Renal biopsy shows **intravascular spindle-shaped vacuoles**. Which of the following is the most likely cause of this patient's symptoms?

Query rewriting:

What is the most probable diagnosis for a 61-year-old male patient with **mottled, reticulated purplish discoloration of the feet**, and **intravascular spindle-shaped vacuoles** in renal biopsy, two weeks after emergency **cardiac catherization with stenting for unstable angina pectoris**, who has …, with laboratory findings of **leukocytosis, eosinophilia, elevated creatinine levels** …?

Query expansion:

Based on the given symptoms and clinical information, possible diseases include:
1.Renal Impairment: Elevated creatinine levels suggest kidney dysfunction, which can be caused by various factors such as **acute kidney injury or chronic kidney disease.**
2.Infectious Diseases: Fever and increased pulse rate may indicate infectious diseases like urinary tract infections or pyelonephritis.
3.Inflammatory Conditions: **Elevated white blood cell count and eosinophil count** suggest the body's response to cholesterol embolization or other inflammatory processes.
4. …

Retrieved Evidence from Textbooks

1. Another name for cholesterol embolism syndrome or atheroembolism is "blue or purple toe syndrome." Atheroembolism that causes digital vascular occlusion causes the syndrome. In this disease, embolisms usually occlude smaller diameter vessels, and hence peripheral pulses are often intact. Therefore, whenever distal gangrene, ulcers, and cyanosis are present, but the pulse is intact, then it is highly suggestive of **blue or purple toe syndrome.**
2. On lab work, usually, signs of systemic inflammation are seen. Complete blood count (CBC) can show evidence of anemia, **leukocytosis, thrombocytopenia**. Hypereosinophilia, hypocomplementemia, elevated amylase, lipase, lactate, elevated LDH, elevated ESR, and high CRP levels are also possible. Urine analysis (UA) can show hematuria, **eosinophiluria,** and proteinuria of varying degrees. The renal function panel shows evidence of **chronic kidney disease or acute kidney injury**, ….
3. …

Based on the provided information, the most likely cause of this patient's symptoms is "**Cholesterol embolization syndrome**" (also known as atheroembolic renal disease or "blue toe syndrome"). Cholesterol embolization occurs when atherosclerotic plaques in larger arteries break off and embolize to smaller vessels, leading to occlusion and tissue damage. In this case, the emboli have affected the kidneys, leading to decreased urinary output and malaise.

Figure 1: Overview of our proposed pipeline. From left to right, we show the Query Augmenter, the Hybrid Textbook Retriever and the LLM Reader through an example of medical question. We have omitted some details from the context for space reasons.

11.0% to 13.2%. Notably, despite its smaller size, medical textbooks outperformed Wikipedia as an external knowledge source, resulting in a performance increase between 9.7% and 12.2%. Furthermore, our human evaluation indicates that our approach effectively reduced hallucinations by 16% in the open-ended QA task when compared to the same baseline. This study shows the effectiveness of our approach to improving model faithfulness.

In summary, the contributions of our research can be articulated as follows: (1) We introduce LLM-AMT, a specialized pipeline tailored specifically for the medical realm. By integrating authoritative medical textbooks as an external knowledge source, we considerably bolster the professionalism and precision of LLMs within this specialized domain. (2) We highlight the intrinsic value of professional domain textbooks in enhancing the expertise of LLMs through rigorous experimentation, thereby shedding light on an intriguing avenue for future research. (3) Our ablation study delves into the optimal configuration for the medical domain, investigating the pivotal roles of the external knowledge retriever, the query augmenter, and the majority voting module within the overarching pipeline.

2 Related Work

In this section, we review the related work in biomedical QA, retrieval-augmented QA and text retrieval.

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).

Retrieval Augmented Generation

When language models tackle tasks that demand supplementary knowledge injection, a retriever can be employed as a knowledge interface. This is due to the fact that language models' training objectives adhere to linguistic rules rather than the goal of constructing a knowledge base. Thus, a retriever that provides external knowledge from a knowledge base, can support more background knowledge to reduce the hallucination from LMs. This process is understood as retrieval augmented generation (Lewis et al. 2020).

The paradigm of retrieval-augmented generation traces its roots to the DrQA framework presented by Chen et al.. In this approach, given a specific question, initial evidence is sourced from Wikipedia using heuristic retrievers like TF-IDF. Subsequently, a neural model is employed to extract answers from this retrieved evidence. In contrast, Dense Passage Retrieval (DPR) (Karpukhin et al. 2020) harnessed the capabilities of pre-trained transformers such as BERT to establish a sophisticated neural retriever and reader, achieving superior performance compared to traditional heuristic models. Following this, Retrieval Augmented Generation (RAG) (Lewis et al. 2020) redefined the methodology by transitioning from answer extraction to generation, thus facilitating the creation of free-form text over mere extractions. In this architecture, both the generator and the retriever undergo joint optimization. Parallel to these advancements, there are works like REALM (Guu et al. 2020) and RETRO (Borgeaud et al. 2022) that aim to enhance language models with retrieval during the pre-training phase. Recently, building on the adept instruction-following capabilities of Large Language Models (LLMs), researchers have delved into the integration of LLMs within the retrieval-augmented generation framework. Notable examples include REPLUG (Shi et al. 2023) and IC-RALM (Ram et al. 2023).

However, most of the past retrieval setting was on general knowledge corpora such as Wikipedia or Web data. In this work, to the best of our knowledge, we present the first study that utilizes a large amount of high-quality medical textbooks as an external knowledge base, integrating multiple retrievers to retrieve medical background knowledge to support the reasoning by the language model.

Neural Text Retrieval

Recent work on Neural Retrieval using bi-encoder architecture has shown significant improvement over traditional retrieval methods such as BM25/TF-IDF. The query and documents are encoded separately by pretrained transformers. And the query–document similarity was then measured by the corresponding embedding distance. Based on the embedding type, the neural retrieval can be classified into, 1) dense retriever, which encodes text into low-dimension dense vectors. Representative works include DPR (Karpukhin et al. 2020), ANCE (Xiong et al. 2020), CoCondenser (Gao and Callan 2021) etc. 2) sparse retriever, which encodes text into high dimension bag-of-words representation. Representative works include DeepImpact (Mallia et al. 2021), uniCOIL (Lin and Ma 2021), SPLADE (Formal, Piwowarski, and Clinchant 2021), etc. 3) late interaction retriever, which keeps token level representation and query–document similarity is aggregated by the similarity of tokens. Representative works include ColBERT (Khattab and Zaharia 2020), COIL (Gao, Dai, and Callan 2021). Although the representation type varies, the model are usually optimized w.r.t the InfoNCE loss:

$$\mathcal{L}(q, D^+, D_1^-, D_2^-, \cdots, D_n^-) = -\log p(D = D^+ \mid Q = q)$$
$$= -\log \frac{\exp(\text{Sim}(q, D^+))}{\exp(\text{Sim}(q, D^+)) + \sum_{i=1}^{n} \exp(\text{Sim}(q, D_i^-))},$$

where D^+ is the positive document to the query q. and D^- is hard negatives or in-batch negatives. And the similarity function $Sim(q, D)$ is usually measured by simple dot product of query and document representation.

In this work, we apply various types of neural retrieval methods to medical textbook retrieval tasks to study their effectiveness in a domain-specific setting beyond the commonly used corpora.

3 LLM-AMT

In this paper, we present LLM-AMT, which is a dedicated process for answering clinical questions. Figure 1 provides an overview of the pipeline, comprising three main steps: (1) The **Query Augmenter** rewrites and expands the input question into a new query. (2) The **Textbook Retriever** collects related evidence from the textbooks using the augmented query. (3) The **LLM Reader** generates the final answer utilizing the retrieved evidence. This task can be formulated as follows. Given a medical open-domain QA benchmark represented by $D = \{(x, y)_i\}$ for $i = 0, 1, 2, \ldots, N$, where x_i denotes a medical question input to the pipeline and y_i stands for the expected output (the correct answer). The question x_i is first rewritten and expanded, resulting in a new query q_i. Subsequently, the retriever sources a set of paragraphs from medical textbooks, which we refer to as the evidence e. The reader then processes the combination $[e, x]$ to predict the output $\hat{y}$.

Query Augmenter

We introduce a query augmenting module designed to enhance queries for more effective retrieval in the medical domain. Our query augmenter consists of two main components: query rewriting and query expansion. For these tasks, we specifically utilize an LLM to achieve both rewriting and expansion of the query.

The process of rewriting the query text involves transforming the terms in the original question into medical terminology. As illustrated in the first part of Figure 1, we utilize a human-written prompt line, *"Rephrase the following question, abstracting the specific symptoms and conditions of the patient into medical terminology"*. This approach preserves key information from the original question, such as details related to *"cardiac catheterization with stenting for unstable angina pectoris"* and *"mottled, reticulated purplish discoloration of the feet"* while facilitating the translation from general description to medical terminology (e.g., *"Leukocyte count 16,400/mm^3; Eosinophils 11%; Creatinine 4.2mg/dL"* is converted into *"leukocytosis, eosinophilia, elevated creatinine levels"*).

Furthermore, to expand the query, we leverage the capabilities of the LLM by instructing it to answer questions using a chain-of-thought approach. We provide it with the directive, *You are a medical doctor, systematically and rigorously reasoning through the following question step by step, and ultimately providing the answers."* This method enables us to uncover more potential directions for retrieval. Examples include queries like *acute kidney injury or chronic kidney disease"* and *"elevated white blood cell count and* *eosinophil count"*, as illustrated in the first part of Figure 1. Through this approach, we also obtain a foundational framework for formulating an answer.

Finally, we generate the augmented query q by concatenating the rewritten query with the expanded query. The Query Augmenter contributes to offering more comprehensive information for the retrieval stage, thereby supporting a more nuanced and effective retrieval process.

Textbook Retrieval Corpus

Medical textbooks, as the epitome of knowledge in the field of 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.

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-words units.

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, Piwowarski, and Clinchant 2021) method. The query and document are encoded separately by BERT, and the MLM layer representation (with dimension 30k) for each token is max aggregated as the text representation. ReLU function was 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, k-NN search is conducted to retrieve the top relevant passages for the given query. For late interaction retrieval, we use the ColBERT setting. For each query token, the token to document similarity is the maximum of the dot products of the query token to all tokens in documents. The query–document similarity score is aggregated by summation over the similarity score of all query tokens.

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" passage as positive passage. 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 evidence retrieval stage, we utilize a fusion of sparse retrieval and dense retrieval as the first-phase recall model. Specifically, we normalize the scores assigned to each document by both the sparse retriever and dense retriever, sum them together, and then select the top k entries. In this paper, k is consistently set to 32. Subsequently, we employ a cross-encoder-based re-ranker to reorder the recalled passages.

LLM Reader

Within the LLM reader, evidence retrieved from medical textbooks provides the foundational context for the input question. This evidence is concatenated with the respective question, subsequently serving as input for the LLM. To ensure the model's understanding and its alignment with our objectives, the following structured prompt is employed:

"You are a medical doctor. Referring to the evidence provided, please examine the subsequent medical query and execute the subsequent tasks:

1. *Critically evaluate the question alongside each potential answer, subsequently determining the correct response.*
2. *Appraise the likelihood of correctness for every option, rating it on a scale from 0 to 10."*

Upon employing this structured prompt, the LLM produces answers utilizing the Chain-of-Thought prompting method. Each retrieved evidence will be used to generate a distinct answer for the same question including the confidence score p_k^i of each option k and the selected option ans^i based on i-th evidence. We then combine N answers using a method based on majority voting, as detailed below:

$$Ans = \arg\max_k \sum_{i=1}^{N} I[ans^i == k] * p_k^i$$

where the confidence scores for each option are summed up and the option with the highest confidence score is selected. In this formula, I represents an indicator function, which outputs a value of 1 if the condition within its brackets is satisfied (true) and 0 otherwise. This ensures that the confidence score is only considered in the summation if the corresponding answer matches the current option k.

4 Experiments

Datasets

We evaluate LLM-AMT on three public 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. Questions in these exams are varied and generally require a deep understanding of related medical concepts from medical textbooks to answer. 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, Umapathi, and Sankarasubbu 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.

Question #	**MedQA-USMLE**	**MedQA-MCMLE**	**MedMCQA**
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

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.

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**, and **PubmedBERT** (Lee et al. 2020; Beltagy, Lo, and Cohan 2019; Yasunaga, Leskovec, and Liang 2022; Gu et al. 2021).

The second group is *Wikipedia-Augmented Models*. These models leverage the knowledge embedded in the Wikipedia corpus to assist in the medical question-answering task. Key models in this category are **Variational ODQA** (Liévin et al. 2023), **Codex 5-shot CoT** (Liévin, Hother, and Winther 2022), and our replicated models using **GPT-3.5-Turbo** and **LLaMA-2-13B** (Touvron et al. 2023) that retrieve evidence from Wikipedia to augment the LLM reader.

Implementation Details

We employed OpenAI's GPT-3.5-Turbo and LLaMA-2-13B as our LLM readers in different experiments. GPT-3.5-Turbo, accessed via its API[1], handled query rewriting during the augmentation phase. In the evidence retrieval stage, SPLADE acted as our sparse retriever, DPR was the dense retriever, and we incorporated a cross-encoder for reranking. The MS-MARCO dataset was 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.

Main Result

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

When compared with closed-book models, our approach, supplemented with textbook knowledge, enhances the GPT-3.5-Turbo's baseline performance by margins of 11.4% and 13.2%. Other domain-specific models, like BioBERT, SciBERT, and PubmedBERT, posted accuracy scores of 36.7%, 39.0%, and 50.3% respectively on the MedQA dataset. While these scores underline the efficacy of domain-centric models, our textbook-augmented method still manages to surpass them. We can also observe a similar trend on the MedMCQA dataset.

Comparing our model to the *GPT-3.5 + Wiki* and *LLaMA + Wiki* setup, while Wikipedia is a rich source, its knowledge may be more generalized and can lack the depth required for specialized fields like medicine. Often, it tends towards providing more layman or popularized insights, making it potentially less valuable for rigorous academic or clinical inquiries. Furthermore, GPT-3.5-turbo might already contain Wikipedia information from its pre-training, possibly leading to minimal performance gains (1.0% increase for MedQA and 1.7% increase for MedMCQA).

In contrast, our method leverages detailed insights from medical textbooks, giving us a clear advantage. We achieve 12.2% increase over *GPT-3.5 + Wiki* for MedQA and 9.7% for MedMCQA. The performance distinction emphasizes the significance of integrating deep, specialized medical knowledge over broad, surface-level information sources.

Component Impact Analysis In our research, we meticulously explored the individual impacts of various components in our designed pipeline, as presented in Table 4. We

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

Method	Retriever	MedQA-USMLE	Med-MCQA
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
Wikipedia-Augmented Model			
Variational ODQA	BM25+DPR	55.0	62.9
Codex 5-shot CoT	BM25	60.2	62.7
LLaMA + Wikipedia	DPR	37.6	39.5
GPT-3.5 + Wikipedia	DPR	52.3	55.6
Textbook-Augmented Model			
LLM-AMT (LLaMA)	HybTextR	42.2	43.8
LLM-AMT (GPT-3.5)	HybTextR	**64.5**	**65.3**

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

demonstrate the insights derived using the example of results from the MedQA-USMLE dataset. The trends we observed are similar in the other two datasets as well.

1. **Textbook Retriever**: Introducing a retrieval mechanism significantly enhanced the performance. This is evident when we observe the leap in accuracy from the baseline GPT-3.5-Turbo score of 51.3% to 58.6% on the MedQA-USMLE dataset upon adding the retriever. It emphasizes the utility of integrating external knowledge bases to extract contextually pertinent information. Moreover, fine-tuning the retriever further bolstered the performance, pushing the accuracy from 62.0% to 64.1% for the "+ retriever + augmented query" configuration.
2. **Query Augmenter**: The efficacy of the query augmenter is highlighted when contrasting the results of "+ retriever" with "+ retriever + augmented query". Specifically, for the MedQA-USMLE dataset, accuracy increased from 58.6% to 62.0% after integrating the augmented query. This underscores the significance of query augmentation in furnishing the retriever with a deeper context and diverse relevant terminologies, thereby enhancing its capability to procure more pertinent evidence.
3. **LLM Reader**: The inclusion of Majority Voting in our architecture further boost the performance. For the MedQA-USMLE dataset, for example, the score grew from 62.0% (with augmented queries) to 62.3% after incorporating majority voting. This shows the LLM reader's ability to consolidate diverse insights from different sources through Majority Voting, which creates a more robust and credible response.

Ablation Study

Here, we perform ablation studies on both the retrieval mechanisms and the majority voting strategy to refine and identify the most optimal configuration specifically tailored for question-answering tasks within the medical domain.

Method	MedQA-USMLE	MedQA-MCMLE	MedMCQA
GPT-3.5-Turbo	51.3	58.2	53.9
+ retriever	58.6	61.2	57.1
+ retriever + majority vote	60.7	62.5	58.8
+ retriever + augmented query	62.0	65.4	63.1
+ retriever + augmented query + majority vote	62.3	66.8	63.9
+ finetuned retriever + augmented query	64.1	68.9	65.2
+ finetuned retriever + augmented query + majority vote	64.5	69.2	65.3

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.

Textbook Retrievers In Table 5, we evaluate the performance impact of various retrieval mechanisms within our pipeline. The late-interaction retriever, particularly the ColBERT method, demonstrates superior accuracy with 58.2% on MedQA-USMLE, 62.4% on MedQA-MCMLE, and 58.1% on MedMCQA in the zero-shot setting, outperforming both the standalone sparse and dense retrievers.

When combining the semantic capabilities of the dense retriever with the precision of the sparse retriever, the system achieves better results. The "Sparse + Dense" combination yields an accuracy of 60.1% on MedQA-USMLE, 64.9% on MedQA-MCMLE, and 58.7% on MedMCQA.

The addition of the reranker further optimizes performance. The "Dense + Rerank" configuration registers scores of 60.6%, 65.4%, and 61.8% across the three datasets, respectively. Notably, the "HybTextR", which integrates sparse, dense, and reranker components, attains the highest performances of 62.0% for MedQA-USMLE, 68.9% for MedQA-MCMLE, and 65.2% for MedMCQA. This consolidated approach demonstrates the benefits of a comprehensive retrieval strategy in the medical domain.

Majority Voting In our LLM reader module related to majority voting, we conducted experiments on different datasets with various top-k settings, distinguishing between cases with and without fine-tuning, as shown in Figure 2. The results indicate that for the fine-tuned retriever, the optimal setting is top-k=2, while for the retriever without fine-tuning, the best performance is achieved with top-k=4. This suggests that fine-tuning significantly alters the optimal top-k configuration for maximizing accuracy.

	Zero-shot			Fine-tuned		
	MedQA-USMLE	**MedQA-MCMLE**	**MedMCQA**	**MedQA-USMLE**	**MedQA-MCMLE**	**MedMCQA**
BM25	55.6	59.7	55.2	–	–	–
Sparse	57.4	60.4	57.5	59.3	62.9	59.6
Dense	59.7	61.0	57.7	60.9	63.8	59.3
ColBERT	58.2	62.4	58.1	61.5	64.1	60.4
Sparse + Dense	60.1	64.9	58.7	62.7	65.5	61.9
Sparse + Rerank	59.5	63.8	59.2	61.3	65.2	62.8
Dense + Rerank	60.6	**65.4**	61.8	63.7	65.3	64.6
HybTextR	**62.0**	64.4	**63.1**	**64.1**	**68.9**	**65.2**

Table 5: Evaluation of Retrieval and Reranking Strategies on the Performance of LLM-AMT

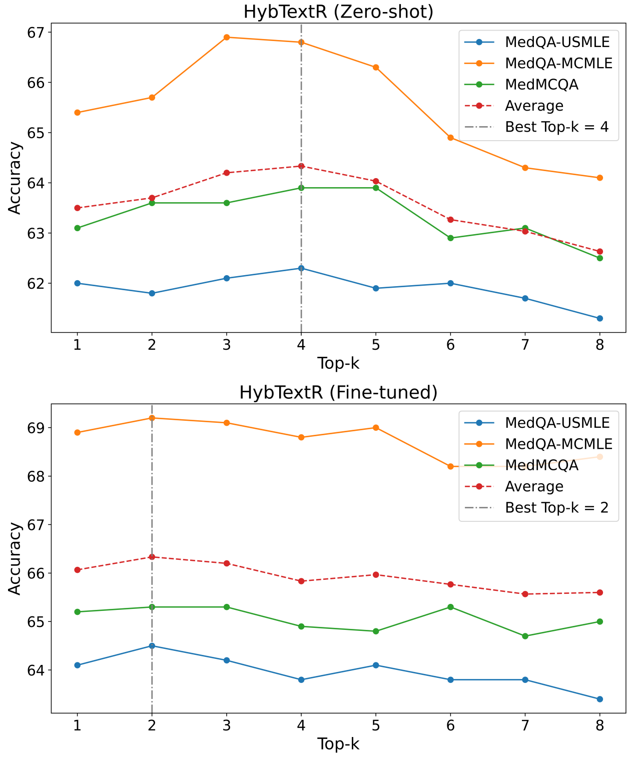


Figure 2: Accuracy of LLM-AMT Under Different Top-k Selections in Majority Voting. GPT-3.5-Turbo was used as the LLM Reader.

Further Discussion

In this section, we provide an in-depth discussion and further assessment of our models, especially their capabilities in handling non-multiple-choice medical QA tasks.

Non-multiple-choice QA Task To emulate the real-world application of models in medical question answering, we randomly selected a diverse subset of 100 questions from the MedQA-USMLE dataset and generated answers directly without accessing the options. For evaluation, individuals with a medical background were enlisted to critically assess the quality of each answer. These answers were categorized into four distinct tiers:

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

Table 6: Evaluation of the Non-multiple-choice Medical Question Answering Task. GPT-3.5 as the LLM Reader.

- **Correct**: Fully accurate and comprehensive.
- **Mostly Correct**: Accurate with minor omissions.
- **Partially Correct**: Contains accurate elements but misses vital details in the response.
- **Wrong**: Largely inaccurate or off-target.

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 6. 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.

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 majority voting mechanisms 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.

References

Beltagy, I.; Lo, K.; and Cohan, A. 2019. SciBERT: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676.*

Borgeaud, S.; Mensch, A.; Hoffmann, J.; Cai, T.; Rutherford, E.; Millican, K.; Van Den Driessche, G. B.; Lespiau, J.-B.; Damoc, B.; Clark, A.; et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, 2206–2240. PMLR.

Chen, D.; Fisch, A.; Weston, J.; and Bordes, A. 2017. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051.*

Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H. W.; Sutton, C.; Gehrmann, S.; et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311.*

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, E.; Wang, X.; Dehghani, M.; Brahma, S.; et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416.*

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

Formal, T.; Piwowarski, B.; and Clinchant, S. 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*, 2288–2292.

Gao, J.; Xiong, C.; Bennett, P.; and Craswell, N. 2022. Neural approaches to conversational information retrieval. *arXiv preprint arXiv:2201.05176.*

Gao, L.; and Callan, J. 2021. Condenser: a pretraining architecture for dense retrieval. *arXiv preprint arXiv:2104.08253.*

Gao, L.; Dai, Z.; and Callan, J. 2021. COIL: Revisit exact lexical match in information retrieval with contextualized inverted list. *arXiv preprint arXiv:2104.07186.*

Gu, Y.; Tinn, R.; Cheng, H.; Lucas, M.; Usuyama, N.; Liu, X.; Naumann, T.; Gao, J.; and Poon, H. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1): 1–23.

Guu, K.; Lee, K.; Tung, Z.; Pasupat, P.; and Chang, M. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, 3929–3938. PMLR.

Izacard, G.; Caron, M.; Hosseini, L.; Riedel, S.; Bojanowski, P.; Joulin, A.; and Grave, E. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118.*

Izacard, G.; Lewis, P.; Lomeli, M.; Hosseini, L.; Petroni, F.; Schick, T.; Dwivedi-Yu, J.; Joulin, A.; Riedel, S.; and Grave, E. 2022. Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299.*

Jin, D.; Pan, E.; Oufattole, N.; Weng, W.-H.; Fang, H.; and Szolovits, P. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14): 6421.

Jin, Q.; Dhingra, B.; Liu, Z.; Cohen, W. W.; and Lu, X. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146.*

Jin, Q.; Yuan, Z.; Xiong, G.; Yu, Q.; Ying, H.; Tan, C.; Chen, M.; Huang, S.; Liu, X.; and Yu, S. 2022. Biomedical question answering: a survey of approaches and challenges. *ACM Computing Surveys (CSUR)*, 55(2): 1–36.

Karpukhin, V.; Oğuz, B.; Min, S.; Lewis, P.; Wu, L.; Edunov, S.; Chen, D.; and Yih, W.-t. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906.*

Khattab, O.; and Zaharia, M. 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*, 39–48.

Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; and Kang, J. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4): 1234–1240.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33: 9459–9474.

Liévin, V.; Hother, C. E.; and Winther, O. 2022. Can large language models reason about medical questions? *arXiv preprint arXiv:2207.08143.*

Liévin, V.; Motzfeldt, A. G.; Jensen, I. R.; and Winther, O. 2023. Variational Open-Domain Question Answering. In *International Conference on Machine Learning*, 20950–20977. PMLR.

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

Luo, R.; Sun, L.; Xia, Y.; Qin, T.; Zhang, S.; Poon, H.; and Liu, T.-Y. 2022. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6): bbac409.

Mallia, A.; Khattab, O.; Suel, T.; and Tonellotto, N. 2021. Learning passage impacts for inverted indexes. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1723–1727.

Mialon, G.; Dessì, R.; Lomeli, M.; Nalmpantis, C.; Pasunuru, R.; Raileanu, R.; Rozière, B.; Schick, T.; Dwivedi-Yu, J.; Celikyilmaz, A.; et al. 2023. Augmented language models: a survey. *arXiv preprint arXiv:2302.07842.*

Pal, A.; Umapathi, L. K.; and Sankarasubbu, M. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning*, 248–260. PMLR.

Peng, B.; Galley, M.; He, P.; Cheng, H.; Xie, Y.; Hu, Y.; Huang, Q.; Liden, L.; Yu, Z.; Chen, W.; et al. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*.

Ram, O.; Levine, Y.; Dalmedigos, I.; Muhlgay, D.; Shashua, A.; Leyton-Brown, K.; and Shoham, Y. 2023. In-context retrieval-augmented language models. *arXiv preprint arXiv:2302.00083*.

Shi, W.; Min, S.; Yasunaga, M.; Seo, M.; James, R.; Lewis, M.; Zettlemoyer, L.; and Yih, W.-t. 2023. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*.

Singhal, K.; Tu, T.; Gottweis, J.; Sayres, R.; Wulczyn, E.; Hou, L.; Clark, K.; Pfohl, S.; Cole-Lewis, H.; Neal, D.; et al. 2023. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; et al. 2022a. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35: 24824–24837.

Wu, B.; Zhang, Z.; Wang, J.; and Zhao, H. 2021. Sentence-aware contrastive learning for open-domain passage retrieval. *arXiv preprint arXiv:2110.07524*.

Xiong, L.; Xiong, C.; Li, Y.; Tang, K.-F.; Liu, J.; Bennett, P.; Ahmed, J.; and Overwijk, A. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*.

Yasunaga, M.; Leskovec, J.; and Liang, P. 2022. Linkbert: Pretraining language models with document links. *arXiv preprint arXiv:2203.15827*.

Zhang, X.; Wu, J.; He, Z.; Liu, X.; and Su, Y. 2018. Medical exam question answering with large-scale reading comprehension. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.