Knowledge Enhancement of Small-Scale Models in Medical Question Answering

Xinbai Li, Shaowen Peng, Shoko Wakamiya, Eiji Aramaki

Submitted to: JMIR Medical Informatics
on: November 03, 2024

Table of Contents

Knowledge Enhancement of Small-Scale Models in Medical Question Answering

Xinbai Li[1]; Shaowen Peng[1] PhD; Shoko Wakamiya[1] PhD; Eiji Aramaki[1] PhD

[1]Nara Institute of Science and Technology Ikoma JP

Corresponding Author:
Eiji Aramaki PhD
Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma, Nara 630-0192, JAPAN
Ikoma
JP

Abstract

Background: Medical question answering (QA) is essential for various medical applications. While small-scale pre-training language models (PLMs) are widely adopted in open-domain QA tasks through fine-tuning with related datasets, applying this approach in the medical domain requires significant and rigorous integration of external knowledge. Knowledge-enhanced small-scale PLMs have been proposed to incorporate knowledge bases (KBs) to improve performance, as KBs contain vast amounts of factual knowledge. Large language models (LLMs) contain a vast amount of knowledge and have attracted significant research interest due to their outstanding natural language processing (NLP) capabilities. KBs and LLMs can provide external knowledge to enhance small-scale models in medical QA.

Objective: KBs consist of structured factual knowledge that must be converted into sentences to align with the input format of PLMs. However, these converted sentences often lack semantic coherence, potentially causing them to deviate from the intrinsic knowledge of KBs. LLMs, on the other hand, can generate natural, semantically rich sentences, but they may also produce irrelevant or inaccurate statements. Retrieval-augmented generation (RAG) paradigm enhances LLMs by retrieving relevant information from an external database before responding. By integrating LLMs and KBs using the RAG paradigm, it is possible to generate statements that combine the factual knowledge of KBs with the semantic richness of LLMs, thereby enhancing the performance of small-scale models. In this paper, we explore a RAG fine-tuning method, RAG-mQA, that combines KBs and LLMs to improve small-scale models in medical QA.

Methods: In the RAG fine-tuning scenario, we adopt medical KBs as an external database to augment the text generation of LLMs, producing statements that integrate medical domain knowledge with semantic knowledge. Specifically, KBs are used to extract medical concepts from the input text, while LLMs are tasked with generating statements based on these extracted concepts. In addition, we introduce two strategies for constructing knowledge: KB-based and LLM-based construction. In the KB-based scenario, we extract medical concepts from the input text using KBs and convert them into sentences by connecting the concepts sequentially. In the LLM-based scenario, we provide the input text to an LLM, which generates relevant statements to answer the question. For downstream QA tasks, the knowledge produced by these three strategies is inserted into the input text to fine-tune a small-scale PLM. F1 and exact match (EM) scores are employed as evaluation metrics for performance comparison. Fine-tuned PLMs without knowledge insertion serve as baselines. Experiments are conducted on two medical QA datasets: emrQA (English) and MedicalQA (Chinese).

Results: RAG-mQA achieved the best results on both datasets. On the MedicalQA dataset, compared to the KB-based and LLM-based enhancement methods, RAG-mQA improved the F1 score by 0.59% and 2.36%, and the EM score by 2.96% and 11.18%, respectively. On the emrQA dataset, the EM score of RAG-mQA exceeded those of the KB-based and LLM-based methods by 4.65% and 7.01%, respectively.

Conclusions: Experimental results demonstrate that RAG fine-tuning method can improve the model performance in medical QA. RAG-mQA achieves greater improvements compared to other knowledge-enhanced methods. Clinical Trial: This study does not involve trial registration.

(JMIR Preprints 03/11/2024:68320)
DOI: https://doi.org/10.2196/preprints.68320

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original Paper

Xinbai Li[†], Shaowen Peng[†], Shoko Wakamiya[†], Eiji Aramaki[†]

[†] Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma, Japan

*Corresponding author:
Eiji Aramaki
Email address: aramaki@is.naist.jp

Knowledge Enhancement of Small-Scale Models in Medical Question Answering

Abstract

Background: Medical question answering (QA) is essential for various medical applications. While small-scale pre-training language models (PLMs) are widely adopted in open-domain QA tasks through fine-tuning with related datasets, applying this approach in the medical domain requires significant and rigorous integration of external knowledge. Knowledge-enhanced small-scale PLMs have been proposed to incorporate knowledge bases (KBs) to improve performance, as KBs contain vast amounts of factual knowledge. Large language models (LLMs) contain a vast amount of knowledge and have attracted significant research interest due to their outstanding natural language processing (NLP) capabilities. KBs and LLMs can provide external knowledge to enhance small-scale models in medical QA.

Objective: KBs consist of structured factual knowledge that must be converted into sentences to align with the input format of PLMs. However, these converted sentences often lack semantic coherence, potentially causing them to deviate from the intrinsic knowledge of KBs. LLMs, on the other hand, can generate natural, semantically rich sentences, but they may also produce irrelevant or inaccurate statements. Retrieval-augmented generation (RAG) paradigm enhances LLMs by retrieving relevant information from an external database before responding. By integrating LLMs and KBs using the RAG paradigm, it is possible to generate statements that combine the factual knowledge of KBs with the semantic richness of LLMs, thereby enhancing the performance of small-scale models. In this paper, we explore a RAG fine-tuning method, RAG-mQA, that combines KBs and LLMs to improve small-scale models in medical QA.

Methods: In the RAG fine-tuning scenario, we adopt medical KBs as an external database to augment the text generation of LLMs, producing statements that integrate medical domain knowledge with semantic knowledge. Specifically, KBs are used to extract medical concepts from the input text, while LLMs are tasked with generating statements based on these extracted concepts. In addition, we introduce two strategies for constructing knowledge: KB-based and LLM-based construction. In the KB-based scenario, we extract medical concepts from the input text using KBs and convert them into sentences by connecting the concepts sequentially. In the LLM-based scenario, we provide the input text to an LLM, which generates relevant statements to answer the question. For downstream QA tasks, the knowledge produced by these three strategies is inserted into the input text to fine-tune a small-scale PLM. F1 and exact match (EM) scores are employed as evaluation metrics for performance comparison. Fine-tuned PLMs without knowledge insertion serve as baselines. Experiments are conducted on two medical QA datasets: emrQA (English) and MedicalQA (Chinese).

Results: RAG-mQA achieved the best results on both datasets. On the MedicalQA dataset, compared to the KB-based and LLM-based enhancement methods, RAG-mQA improved the F1 score by 0.59% and 2.36%, and the EM score by 2.96% and 11.18%, respectively. On the emrQA dataset, the

EM score of RAG-mQA exceeded those of the KB-based and LLM-based methods by 4.65% and 7.01%, respectively.
Conclusions: Experimental results demonstrate that RAG fine-tuning method can improve the model performance in medical QA. RAG-mQA achieves greater improvements compared to other knowledge-enhanced methods.

Keywords: medical question-answering; retrieval-augmented generation; knowledge bases; large language models

Introduction

Medical artificial intelligence (AI) has garnered significant research interest, with various applications such as medical recommendation systems, clinical decision-making support systems, and medical chatbots—based on medical question-answering (QA)—playing an increasingly critical role [1–3]. Small-scale pre-training language models (PLMs) have been widely applied in open-domain QA tasks. Compared to open-domain QA, medical QA involves more complex technical terminology and requires models to have a deeper understanding of medical knowledge [4,5]. Small-scale models fine-tuned on medical QA datasets for downstream tasks often lack comprehensive medical knowledge.

Knowledge-enhanced methods have been proposed to inject external knowledge into small-scale models to improve their performance [6]. For example, given the question “Was the patient ever prescribed a non-steroidal anti-inflammatory agent?” and the electronic medical record (EMR) stating, “The patient was treated with aspirin,” the semantics of the EMR alone do not provide an intuitive answer. However, the medical knowledge that “Aspirin is a non-steroidal anti-inflammatory agent” provides the necessary information to answer the question.

Generally, knowledge-enhanced methods inject knowledge in pre-training stage and fine-tune models for downstream tasks [7,8]. For instance, M-cERNIE [9] enriches the representation of each medical entity in pre-training stage to improve model performance. Pre-training requires a substantial amount of computational resources [10]. For specific downstream tasks like medical QA, this paper explores knowledge-enhanced fine-tuning methods in resource-constrained environments.

Medical concepts and their relations constitute medical knowledge bases (KBs), which are usually utilized as additional knowledge sources due to their comprehensive and specialized information. Structured KBs need to be converted into text to fit the input format of PLMs. Existing conversion methods construct statements by sequentially connecting medical terms and relations [11]. For a single relation between two concepts, the conversion methods are direct and effective. However, in complex cases involving multiple terms and relations, these methods struggle to preserve the relational descriptions from KBs, leading to statements that deviate from the original knowledge and may produce ambiguous sentences.

Large language models (LLMs) have been used across a range of natural language processing (NLP) tasks [12–14]. Retrieval-augmented generation (RAG) paradigm aims to augment LLMs by retrieving related information before generating [15,16]. Although RAG-based LLMs can be applied to medical QA tasks through in-context learning (ICL), their generative responses often include extraneous information beyond the answers, resulting in outputs mixed with noise. Furthermore, fine-tuning LLMs also consumes a significant amount of computational resources. Instead of enhancing LLMs using RAG paradigm, this paper explores the knowledge enhancement of small-scale models under limited computational resources.

LLMs encompass rich semantic knowledge and can generate natural statements that conform to the input format of models. However, the generated statements may not always align with factual knowledge and may even produce sentences unrelated to the topic. Therefore, it is essential to combine KBs and LLMs to construct accurate and reliable knowledge.

In this paper, we propose a RAG fine-tuning method, RAG-mQA, to enhance small-scale PLMs by incorporating concise, accurate and complete knowledge. In the RAG-based knowledge construction scenario, KBs are utilized as an external database to retrieve relevant medical information, while LLMs are tasked with generating clear and precise statements based on the retrieved information. These generated statements are then inserted into the input text as external knowledge to fine-tune small-scale PLMs. Additionally, we introduce two other strategies for knowledge construction. In the KB-based construction strategy, text converted from KBs is inserted into the input. In the LLM-based construction strategy, LLMs generate useful knowledge to assist in answering questions. These strategies are compared on medical QA tasks to demonstrate the effectiveness of different knowledge sources.

The main contributions of this work are listed as follows:

1. Unlike existing knowledge-enhanced PLMs, we propose a RAG fine-tuning method to enhance medical QA tasks under limited computational resources.
2. We empirically discover the superiority of KB-based over LLM-based methods, despite LLM's powerful language generation capability and vast real-world knowledge, demonstrating the importance of domain-specific knowledge in medical QA domain.
3. Experimental results demonstrate that knowledge-enhanced methods improve the performance of small-scale models in medical QA tasks. RAG-mQA achieves greater improvements compared to KB-based and LLM-based methods.

Related Work

Knowledge-Enhanced PLMs

Integrating factual knowledge into pre-trained models forms the basis of the knowledge-enhanced paradigm. Methods such as ERNIE-THU [17] and KnowBERT [18] independently train static entity embeddings sourced from KBs and PLMs. WKLM [19] enhances model training by replacing entity names in the text with other entities of the same type. ERNIE-Baidu [6] employs phrase-level and entity-level masking techniques to incorporate factual information. K-Adapter [20] combines factual and linguistic knowledge through the use of neural adapters. Approaches like K-BERT [11] and CoLAKE [21] build graphs that merge sentences with knowledge graphs (KGs), capturing semantic knowledge via graph-based learning methods. KEPLER [22] and KLMo [7] use attentive fusion modules to integrate pre-trained embeddings from KBs into PLMs. Although these knowledge-enhanced PLMs are effective, they require significant computational resources during pre-training. Therefore, for downstream tasks such as medical QA, knowledge-injected fine-tuning is a more practical approach than knowledge-enhanced pre-training.

Knowledge-Injected Fine-Tuning

While knowledge-enhanced pre-training paradigms involve adjusting the entire set of model weights, knowledge-injected fine-tuning focuses on optimizing specific model parameters to better suit the nuances and requirements of downstream tasks. For instance, the researchers [10] fine-tune PLMs using emotion vocabulary as a knowledge source, enhancing emotion recognition by improving the embeddings of emotional words. Similarly, the paper [23] predicts KBs information based on input text to improve the model's semantic representation for dialogue generation. However, these knowledge-injected fine-tuning methods rely solely on a KB as an external knowledge source, which

may limit the breadth of knowledge incorporated into the model.

Knowledge-Enhanced Medical QA

Knowledge-enhanced medical QA aims to improve the performance of medical QA systems by incorporating medical knowledge into models. [24] explore the integration of disease-specific knowledge with PLMs such as BERT, BioBERT [25] and ClinicalBERT [26] to enhance tasks like consumer health QA, medical language inference, and disease name recognition. BioLORD [27] is designed to generate meaningful representations for clinical sentences and biomedical concepts by grounding these representations using definitions and descriptions from a multi-relational biomedical KB, resulting in more semantically and ontologically aligned representations. M-cERNIE [9] enriches PLMs by introducing medical concepts from a medical KB. Similar to knowledge-injected fine-tuning, existing knowledge-enhanced medical QA methods typically rely on a single KB and lack the incorporation of knowledge from multiple sources.

Retrieval-Augmented Generation

Retrieval-augmented generation methods commonly retrieve external databases based on the input and then enable LLMs to generate text conditioning on the retrieved information. KnowledGPT [28] integrates LLMs with various KBs to address issues like completeness, timeliness, faithfulness, and adaptability in LLMs. It facilitates both the retrieval and storage of knowledge by employing program-of-thought prompting to generate search language for KBs in code format with pre-defined functions, and allows users to store knowledge in personalized KBs. search module. REINA [29] enhances NLP tasks by retrieving and concatenating the most similar labeled training instances with the input text, reducing computational costs associated with large-scale external retrieval. Selfmem [30] RAG by leveraging the model's own outputs as an unbounded memory pool, thus overcoming limitations of traditional fixed-corpus retrieval methods. RA-DIT [31] is a lightweight fine-tuning methodology that retrofits any LLM with retrieval capabilities without requiring expensive pre-training modifications or suboptimal post-hoc integration. UPRISE [32] is also a lightweight and versatile retriever that automatically selects prompts for any zero-shot task input, eliminating the need for model-specific fine-tuning or task-specific prompt engineering. Promptagator [33] leverages LLMs to generate queries based on a few examples, creating task-specific retrievers without relying on large datasets. Self-RAG [34] allows the model to retrieve passages on-demand and reflect on both the retrieved content and its own generated responses using special reflection tokens. Existing RAG methods aim to enhance the performance of LLMs in various NLP tasks by combining multiple data sources. However, employing LLMs consumes significant computational resources, and their generative outputs can make the response content uncontrollable. Therefore, in contrast to existing RAG research, this paper explores a RAG fine-tuning method that operates under limited computational resources to improve the performance of small-scale models in medical QA tasks.

Methods

In the knowledge-enhanced medical QA scenario, the key is to extract suitable knowledge that can effectively boost model performance in answering questions. Generally, this scenario involves two steps: knowledge construction and model fine-tuning. We propose a RAG fine-tuning method to combine the knowledge of medical KBs and LLMs to enhance small-scale PLMs. Additionally, we introduce two knowledge construction strategies: KB-based and LLM-based strategies. The same fine-tuning process is applied across all three construction strategies.

Given a paragraph S and a question Q, our objective is to extract knowledge K based on a KB B and an LLM L.

RAG-Based Knowledge Construction

Medical KBs contain large-scale structured information that provides credible external knowledge to support medical QA tasks. The Unified Medical Language System (UMLS) [35] encompasses over four million medical concepts and their interrelations. We utilize UMLS as the medical KB B to extract relevant concepts. Additionally, to evaluate the performance of our method in different language contexts, the Chinese Medical Knowledge Graph (CMeKG)[1] is employed for Chinese medical QA tasks.

We utilize SciSpaCy [36] toolkit to extract medical concepts from English QA pair S and Q. Similarly, CMeKG provides a toolkit to extract medical concepts from Chinese QA pairs. As shown in Figure 1, all concepts extracted from a QA pair are connected sequentially and converted into a sentence.

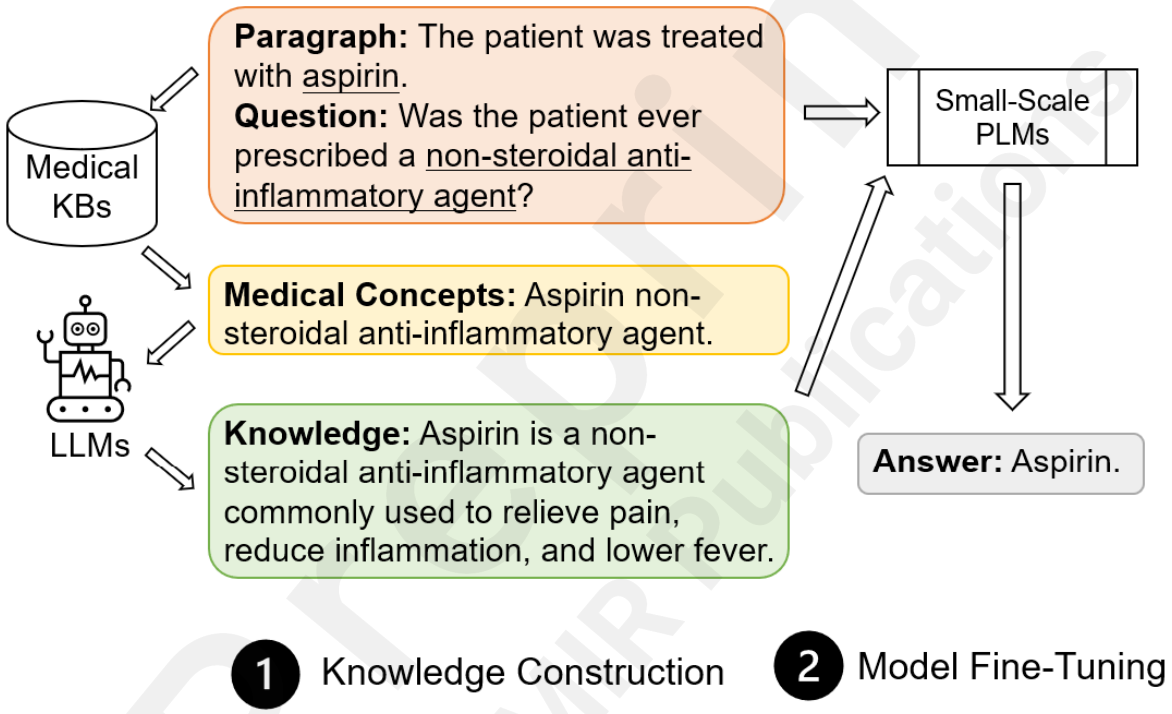


Figure 1. An Example of RAG-Based Model Fine-Tuning. The underlined words in the question and paragraph represent extracted medical concepts.

The sentences converted from the KB are often disordered and incomplete. For example, the sentence in Figure 1, “Aspirin non-steroid anti-inflammatory agent,” is incomplete. In contrast, LLMs possess vast semantic knowledge and excel in generation tasks, demonstrating the ability to produce high-quality natural language text. The LLaMa2 [37] series model is a widely used open-source LLM. In this paper, we evaluate the 13B chat model. The medical concepts extracted from the KB are provided to the LLM, which is tasked with generating natural sentences based on these concepts. The template for knowledge generation prompt is shown in Figure 2. The sentence generated by the LLM L is the knowledge K.

[1] https://github.com/king-yyf/CMeKG_tools

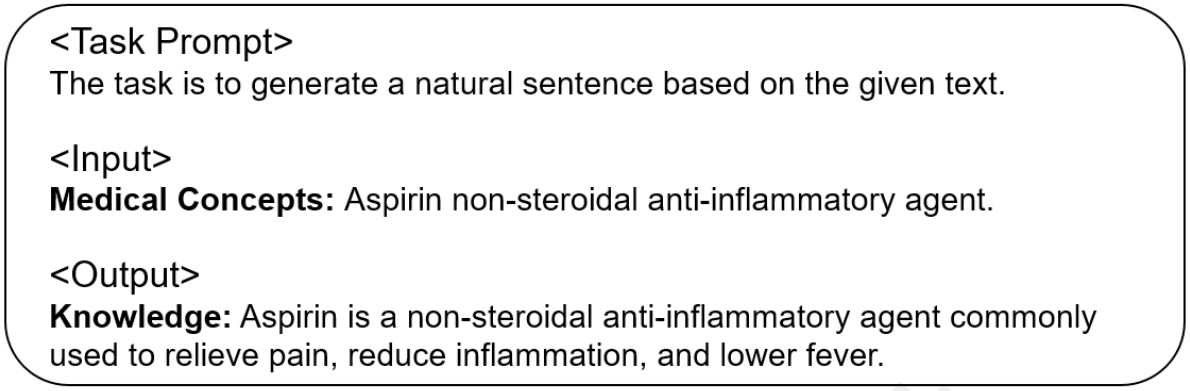


Figure 2. The Prompt Template of RAG-Based Knowledge Construction.

KB-Based and LLM-Based Knowledge Construction

For comparison, we introduce KB-based and LLM-based knowledge construction, respectively. The medical concepts extracted from the KB, when connected sequentially, form a body of knowledge, denoted as K^B. In the LLM-based construction, model L is tasked with generating related knowledge K^L to facilitate answering the question. The prompt template is given in Figure 3.

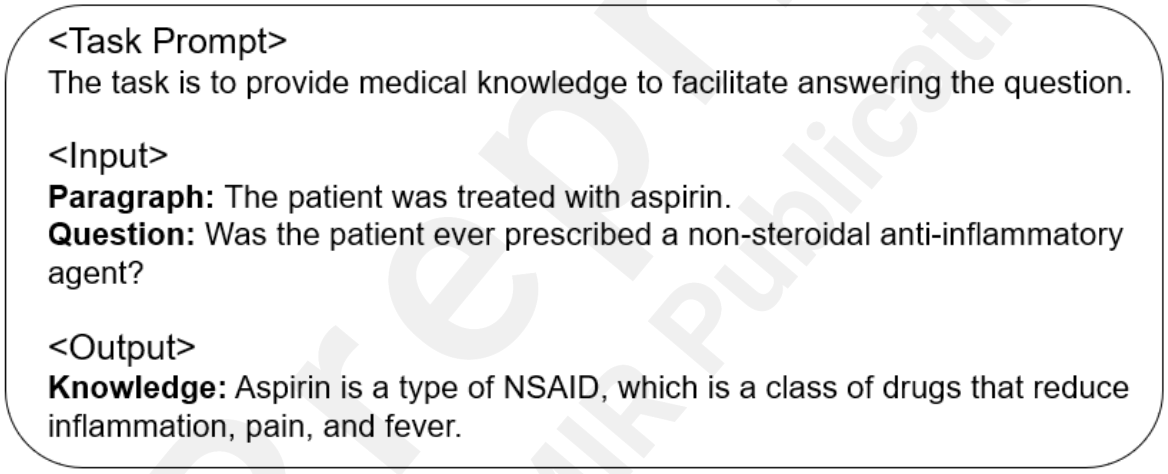


Figure 3. The Prompt Template of LLM-Based Knowledge Construction.

Model Fine-Tuning

Constructed knowledge is inserted into input text to train models. We utilize the same fine-tuning process to explore the knowledge enhancement effects of different knowledge construction strategies.

We adopt the widely used generative PLM T5 [38], which converts all text-based language tasks into a text-to-text format, as the base model. Medical QA is cast as feeding the model paragraphs and questions as input text and training it to generate answers. Each training data includes an input text I and an answer A. The input text is concatenated as:

$$I = Concat(S, Q, K) \quad (1)$$

where I represent RAG-enhanced input text, $Concat(\cdot,\cdot,\cdot)$ denotes sequence concatenation, $I^B = Concat(S, Q, K^B)$ and $I^L = Concat(S, Q, K^L)$ represent KB-enhanced and LLM-enhanced input text, respectively.

The fine-tuning target is to minimize the difference between the generated sequence and the target sequence. Take the RAG-based enhancing method as an example, for a medical dataset

$D=\{(I_1,A_1),(I_2,A_2),\ldots,(I_d,A_d)\}$ which contains d samples, the fine-tuning target is defined as:

$$\theta^{*}=\arg\min_{\theta}\sum_{i=1}^{d}L\left(A_i,P\left(A_i|I_i;\theta\right)\right) \quad (2)$$

where θ represents the pre-trained model parameters, L represents the cross-entropy loss, $P\left(A_i\middle|I_i;\theta\right)$ denotes the conditional probability of generating the target sequence A_i, and I_i is replaced with I_i^B and I_i^L in the KB-based and LLM-based enhancement scenario, respectively.

Experiments

In this section, we provide specific experimental settings including dataset information, baseline methods, evaluation metrics, and implementation details.

Dataset

We evaluate our method on two datasets, emrQA and MedicalQA. The statistics of the above datasets are shown in Table 1.

The emrQA [39] is a community-shared English clinical QA dataset. EMRs contain a record of a patient's health information in the form of unstructured clinical notes (e.g., progress notes and discharge summaries) as well as structured vocabularies. Medical QA tasks based on EMRs are to answer questions posed by physicians against patient EMRs, and the answers are included in the EMRs.

The MedicalQA[2] is a Chinese medical QA dataset where the questions are commonly asked by patients, and the answers can be extracted from the given paragraph. Questions can be divided into nine categories: internal medicine, surgery, gynecology and obstetrics, pediatrics, dermatology, venereal diseases, facial features, traditional Chinese medicine, and infectious diseases.

Table 1. Statistics of datasets.

Dataset	Language	# Train	# Dev	# Test
emrQA	EN	133,589	21,666	19,401
MedicalQA	ZH	15,600	1,951	1,951

Baseline and Evaluation Metrics

The T5 model has been released with pre-trained weights as T5-small, T5-base, and T5-large. We employ T5-small as the PLM to explore the knowledge enhancement effects. The results on the emrQA dataset for BERT [40] and its knowledge-enhanced variant M-cERNIE, which incorporates knowledge during the pre-training stage, are reported for comparison. We adopt the T5-small model as the baseline for both datasets. Our RAG-based enhancement method is named as RAG-mQA. The KB-based and LLM-based enhancement methods for the T5 are named B-mQA and L-mQA, respectively.

The details are provided as follows:
BERT: BERT-base model.
M-cERNIE: Entity-enriched BERT-base model.
T5: T5-small model.
B-mQA: KB-enhanced T5-small model.
L-mQA: LLM-enhanced T5-small model.

[2] https://www.scidb.cn/en/detail?dataSetId=c866b22e95b64928bfa151eecb6cdfbe

RAG-mQA: RAG-enhanced T5-small model.

For emrQA dataset, we utilize exact match (EM) and F1 for evaluation. For MedicalQA dataset, only the F1 is used for evaluation since the answers span multiple sentences.

Ablation

To further explore the impact of knowledge enhancement, we apply knowledge insertion into the LLM to compare the differences in knowledge effects between the small-scale PLM and the LLM.

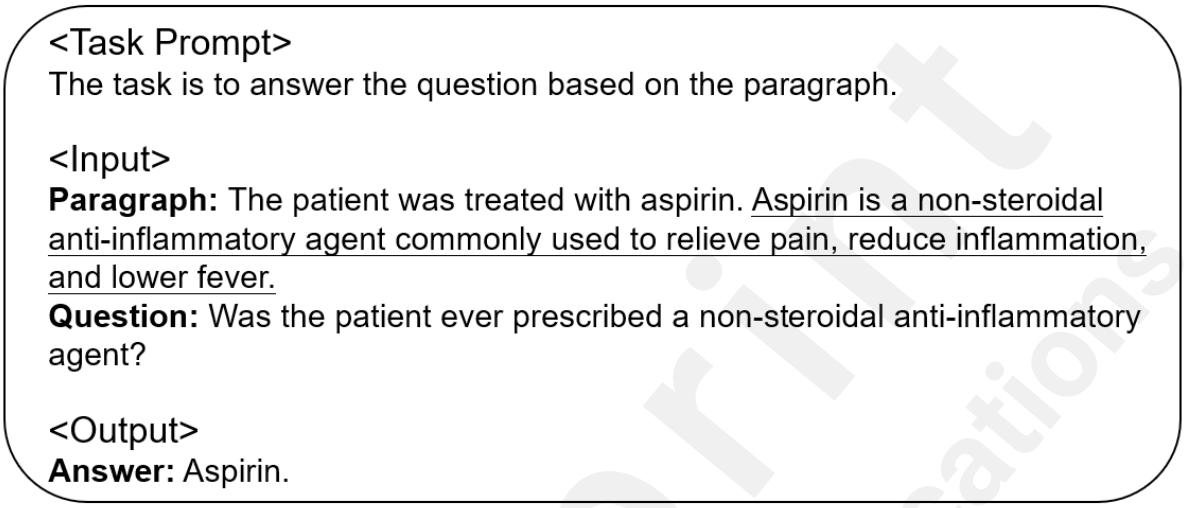


Figure 4. The Prompt Template of QA Task. The underlined sentence represents the inserted knowledge in knowledge enhancement scenarios.

We employ the LLaMa2-13B-chat model to answer questions, referred to as the LLM-only approach. Paragraphs and questions are input into the model to generate answers. In knowledge-enhanced scenarios, constructed knowledge is inserted into the paragraph to assist in answering questions. The QA prompt template is shown in Figure 4, where sentences with underlines represent the inserted knowledge. Three knowledge construction strategies are utilized and named as RAG-LLM, B-LLM, and L-LLM, respectively. The details are provided as follows:
LLM-only: LLaMa2-13B-chat model.
B-LLM: KB-enhanced LLaMa2-13B-chat model.
L-LLM: LLM-enhanced LLaMa2-13B-chat model.
RAG-LLM: RAG-enhanced LLaMa2-13B-chat model.

Implementation Details

The temperature, top k, and top p of LLaMa2 are set as 0.5, 5, and 0.5, respectively. The optimizer of fine-tuning is ADMA, and the learning rate is 5×10^{-5}. F1 score is utilized as the primary metric for early stopping, and patience is set to 3.

Results

The results on the emrQA and MedicalQA datasets are shown in Table 2 and 3, respectively. RAG-mQA achieves the best results on both datasets. As shown in Table 2, the F1 score of RAG-mQA surpasses that of T5 and BERT by 6.37% and 18.12%, respectively. Moreover, the EM score of RAG-mQA exceeds that of T5 and BERT by 9.35% and 23.61%, respectively. The F1 scores of B-mQA, L-mQA, and M-cERNIE also exceeds that of T5 and BERT. Similarly, the results in Table 3 show that the F1 scores of the knowledge-enhanced methods—RAG-mQA, B-mQA, and L-mQA—are higher than that of T5 by 14.83%, 10.18%, and 7.82%, respectively, further supporting the notion

that knowledge insertion can effectively improve the performance of medical QA models. The EM score of L-mQA is lower than that of T5. This suggests that although LLM-based knowledge construction can provide richer information, leading to an improvement in the F1 score, excessive irrelevant information introduces noise, resulting in a decrease in the EM score.

Table 2. Results on the emrQA.

Model	F1	EM
BERT	72.13	65.81
M-cERNIE	87.61	77.91
T5	83.88	80.07
B-mQA	89.66	86.46
L-mQA	87.89	78.24
RAG-mQA	**90.25**	**89.42**
LLM-only	58.38	37.15
B-LLM	60.09	38.43
L-LLM	60.63	38.29
RAG-LLM	60.88	38.33

Table 3. Results on the MedicalQA.

Model	F1
T5	27.51
B-mQA	37.69
L-mQA	35.33
RAG-mQA	**42.34**
LLM-only	18.02
B-LLM	18.46
L-LLM	18.89
RAG-LLM	19.11

We insert knowledge during the fine-tuning stage, whereas M-cERNIE incorporates medical entities during the pre-training stage. As shown in Table 2, the F1 and EM scores of RAG-mQA exceed those of M-cERNIE by 2.64% and 11.51%, respectively. Additionally, the F1 and EM scores of B-mQA surpass those of M-cERNIE by 2.05% and 8.55%, while the scores of L-mQA exceed those of M-cERNIE by 0.28% and 0.33%. This demonstrates that our knowledge-enhanced methods can achieve model improvements with minimal computational resources.

RAG-mQA outperforms both B-mQA and L-mQA on both datasets. As shown in Table 2, the F1 score of RAG-mQA is higher than that of B-mQA and L-mQA by 0.59% and 2.36%, respectively. The EM score of RAG-mQA exceeds that of B-mQA and L-mQA by 2.96% and 11.18%, respectively. In Table 3, the F1 score of RAG-mQA also surpasses that of B-mQA and L-mQA by 4.65% and 7.01%. This suggests that the RAG-based knowledge insertion method constructs accurate and complete knowledge which can more effectively enhance model performance. The results of L-mQA are lower than those of B-mQA on both datasets. This indicates that in knowledge-inserted fine-tuning, concise domain-specific knowledge is more meaningful than abundant semantic knowledge.

In ablation experiments, the results of the LLM are significantly lower than those of the fine-tuned small-scale models. On the emrQA dataset, the F1 and EM scores of RAG-LLM are lower than those

of RAG-mQA by 29.37% and 51.09%, respectively. On the MedicalQA dataset, the F1 score of RAG-LLM is lower than that of RAG-mQA by 23.23%. LLMs demonstrate strong performance in generative tasks through ICL, but their responses often contain content unrelated to the answer, which in turn affects their performance in tasks with high accuracy requirements, such as the extractive QA tasks evaluated in this paper. The performance of RAG-LLM is the best among the four knowledge-enhanced LLM methods. The results of all these methods are higher than those of the LLM-only approach, but the improvements are limited. These findings indicate that the RAG-based knowledge construction method can also enhance LLMs; however, the accuracy and completeness of the inserted knowledge have a much less significant impact on the ICL of LLMs compared to the effect on the fine-tuning of small-scale models.

Discussion

We present three examples from the emrQA and MedicalQA datasets in Table 4 and 5, respectively. Overall, the knowledge constructed by RAG is clear and straightforward while also containing rich information. In contrast, the knowledge constructed by the KB is concise but often has ambiguous and fragmentary semantics. The knowledge generated by the LLM typically consists of rich information; however, much of it is irrelevant to the QA topic.

Table 4. Examples from the emrQA.

ID	Question	Paragraph	Ground Truth	Knowledge	
1	What is has been given for treatment of her severe frontal headaches with scintillations?	... no relief with Tylenol, Aspirin or Fioricet. She also stated ...	Aspirin	RAG-mQA	The patient was treated for severe frontal scintillations with Tylenol, aspirin, and Fioricet.
				B-mQA	Treatment severe frontal scintillations Tylenol Aspirin Fioricet.
				L-mQA	Severe frontal headaches with scintillations can be a symptom of a variety of conditions, including migraines, sinusitis, and cluster headaches.
2	Why does the patient take Lasix?	... the patient's pulmonary edema was treated with Lasix ...	Pulmonary edema	RAG-mQA	The patient's pulmonary edema was managed with treatment using Lasix.
				B-mQA	Patient's pulmonary edema treated with Lasix.
				L-mQA	Lasix (furosemide) is a medication commonly used to treat fluid retention (edema) and high blood pressure.
3	What is the patient's current dosage does the patient take of her blood?	... afebrile throughout her hospital course and required one unit of ...	One unit	RAG-mQA	The patient's blood tests were stable throughout their hospital course, and they remained afebrile during treatment.
				B-mQA	Patient's dose patient blood afebrile hospital course.
				L-mQA	The patient's current dose of blood is one unit of packed red blood cells (PRBCs).

As shown in Example 1 in Table 4, the knowledge from RAG clearly states, "the patient was treated

with aspirin," while the knowledge from the KB is unclear. The knowledge constructed by RAG and the KB directly provides the information "treatment aspirin" to answer the question, whereas the knowledge constructed by the LLM deviates from the key point of the question. Although the sentence generated by the LLM contains rich knowledge in Example 2, the correlation between this knowledge and the question is weak. In comparison, the knowledge provided by RAG and the KB is more concise and intuitive. In Example 3, the key aspect of the question revolves around the dosage. The KB fails to provide additional information; since RAG is generated based on the KB, it similarly cannot provide relevant details, while the LLM accurately delivers the critical answer.

Table 5. Examples from the MedicalQA.

ID	Question	Paragraph	Ground Truth	Knowledge	
1	? What are the symptoms and dangers of low blood lipids?	... ... Low blood lipid levels can easily cause symptoms such as dizziness and headaches. The dangers include triggering a series of complications, such as abdominal infections and menstrual irregularities. Extremely low blood lipid levels can also lead to reduced liver function...	Low blood lipid levels can easily cause symptoms such as dizziness and headaches. The dangers include triggering a series of complications, such as abdominal infections and menstrual irregularities.	RAG-mQA	Low blood lipid levels can easily cause symptoms such as dizziness, headaches, infections, and menstrual irregularities, while endocrine disorders may present with emotional instability.
				B-mQA	Low blood lipids clinically manifest as dizziness, low blood lipids clinically manifest as headaches, low blood lipids clinically manifest as infections, low blood lipids clinically manifest as menstrual irregularities, and endocrine disorders clinically manifest as emotional instability.
				L-mQA	The symptoms and dangers of low blood lipids include dizziness, headaches, abdominal infections, menstrual irregularities, reduced liver function, ascites, skin swelling caused by ascites, abdominal infections, life-threatening conditions, endocrine disorders, emotional instability, and early menopause.
2	? What is the impact of hyperkalemia on the	... ...	It may cause symptoms of heart failure.	RAG-mQA	The clinical manifestations of hyperkalemia include excitement and arrhythmias.
				B-mQA	Hyperkalemia clinically

	myocardium?	It may cause symptoms of heart failure. The effect of hyperkalemia on myocardial cells is initially excitatory, followed by inhibition. In the early stages, hyperkalemia has an excitatory effect on the heart, increasing myocardial sensitivity, and various types of rapid arrhythmias may also occur...			manifests as excitability, and hyperkalemia clinically manifests as arrhythmias.
				L-mQA	□□□ The effect of hyperkalemia on the myocardium is initially excitatory, followed by inhibition. In the early stages, hyperkalemia excites the heart, increasing the sensitivity of myocardial cells, which leads to a rise in various rapid arrhythmias.
3	□□ hcg □□□? How much does a blood HCG test cost?	□HCG□□□□□□□□□80□-100□□□□hcg□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□…… The cost of a blood HCG test generally ranges from about 80 to 100 yuan. HCG, or human chorionic gonadotropin, is currently the most accurate method for early pregnancy testing...	□HCG□□□□□□□□□80□-100□□□□ The cost of a blood HCG test generally ranges from about 80 to 100 yuan.	RAG-mQA	(Empty)
				B-mQA	(Empty)
				L-mQA	HCG□□□20□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□80□100□□□ The HCG test is one of the most accurate methods for early pregnancy detection and can be conducted using either urine or blood. Urine testing is relatively inexpensive, costing about 20 yuan, but it has a certain degree of inaccuracy. Blood testing, on the other hand, is more expensive, ranging from about 80 to 100 yuan, with slight price variations depending on the reagents used by different hospitals.

Examples from the MedicalQA dataset are shown in Table 5. The knowledge constructed by RAG, KB, and LLM is generally semantically similar in Example 1. The knowledge provided by RAG directly answers the question, while the knowledge from the KB contains a significant amount of repetitive expressions, such as “low blood lipids clinically manifest as …”. The statement generated by the LLM is more natural than that of the KB but contains excessive information. In Example 2, all the knowledge deviates from the key points of the answer. The question in Example 3 pertains to pricing. It is evident that neither the KB nor RAG can provide the relevant information, whereas the LLM extracts the key points of the answer but includes a significant amount of unnecessary details.

In summary, RAG provides clear and concise knowledge. The knowledge constructed by the KB is

succinct but may sometimes lack semantic clarity. The LLM, on the other hand, generates statements with clear semantics and rich information; however, the abundance of information can occasionally introduce knowledge noise (KN), which may affect the model's predictions. Notably, for questions related to numerical values, the LLM is able to extract key information, whereas the KB fails to provide relevant details.

Limitations

This work proposed a knowledge-enhanced medical QA method based on the KB and LLM. We evaluated our RAG fine-tuning method, RAG-mQA, and compared the effectiveness of different knowledge enhancement methods. However, there are still limitations to this work. Although RAG can generate semantically clear knowledge, the extraneous information it produces can lead to KN, which affects the model's predictions. Additionally, we only explored these simple knowledge insertion methods during the fine-tuning stage and have not yet considered knowledge insertion during the pre-training phase or modifications to the model architecture to accommodate knowledge insertion.

Conclusions

In this paper, we propose a RAG fine-tuning method and compare different knowledge-enhanced methods on medical QA tasks. We employ a medical KB and an LLM to extract concepts and generate knowledge. Experimental results demonstrate that the RAG fine-tuning method significantly improves the model performance, and achieves greater enhancements than both KB-based and LLM-based enhancement methods. LLMs perform well in generative tasks, but redundant information can introduce noise, resulting in performance that is inferior to that of KBs in knowledge enhancement.

In the future, we will explore conversion methods to transform KBs into natural sentences and selection methods to identify relevant knowledge from the rich information generated by LLMs. Furthermore, we will investigate modifications to the model architecture to accommodate knowledge insertion.

Acknowledgements

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

Conflicts of Interest

The authors declare no conflicts of interest.

Abbreviations

QA: Question Answering
PLMs: Pre-training Language Models
LLMs: Large Language Models
RAG: Retrieval-Augmented Generation
AI: Artificial Intelligence
NLP: Natural Language Processing
Electronic Medical Records: EMRs
ICL: In-Context Learning

References

1. Mani V, Kavitha C, Band SS, Mosavi A, Hollins P, Palanisamy S. A Recommendation System

Based on AI for Storing Block Data in the Electronic Health Repository. *Front Public Health*. 2022;9. doi:10.3389/fpubh.2021.831404
2. Hsu IC, Yu JD. A medical Chatbot using machine learning and natural language understanding. Multimed Tools Appl. 2022;81(17):23777-23799. doi:10.1007/s11042-022-12820-4
3. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. Npj Digit Med. 2020;3(1):1-10. doi:10.1038/s41746-020-0221-y
4. Wang X, Wang J, Xu B, Lin H, Zhang B, Yang Z. Exploiting Intersentence Information for Better Question-Driven Abstractive Summarization: Algorithm Development and Validation. JMIR Med Inform 2022;10(8): e38052. URL: https://medinform.jmir.org/2022/8/e38052. DOI: 10.2196/38052
5. Wang H, Du H, Qi G, Chen H, Hu W, Chen Z. Construction of a Linked Data Set of COVID-19 Knowledge Graphs: Development and Applications. JMIR Med Inform 2022;10(5): e37215. URL: https://medinform.jmir.org/2022/5/e37215. DOI: 10.2196/37215
6. Sun Y, Wang S, Li Y, et al. ERNIE: Enhanced Representation through Knowledge Integration. Published online April 19, 2019. doi:10.48550/arXiv.1904.09223
7. He L, Zheng S, Yang T, Zhang F. KLMo: Knowledge Graph Enhanced Pretrained Language Model with Fine-Grained Relationships. In: Moens MF, Huang X, Specia L, Yih SW tau, eds. Findings of the Association for Computational Linguistics: EMNLP 2021. Association for Computational Linguistics; 2021:4536-4542. doi:10.18653/v1/2021.findings-emnlp.384
8. Hayashi H, Hu Z, Xiong C, Neubig G. Latent Relation Language Models. Proc AAAI Conf Artif Intell. 2020;34(05):7911-7918. doi:10.1609/aaai.v34i05.6298
9. Rawat BPS, Weng WH, Min SY, Raghavan P, Szolovits P. Entity-Enriched Neural Models for Clinical Question Answering. Published online February 19, 2021. Accessed June 2, 2024. http://arxiv.org/abs/2005.06587
10. Zhu Z, Mao K. Knowledge-based BERT word embedding fine-tuning for emotion recognition. Neurocomputing. 2023;552:126488. doi:10.1016/j.neucom.2023.126488
11. Liu W, Zhou P, Zhao Z, et al. K-BERT: Enabling Language Representation with Knowledge Graph. Proc AAAI Conf Artif Intell. 2020;34(03):2901-2908. doi:10.1609/aaai.v34i03.5681
12. Sun X, Dong L, Li X, et al. Pushing the limits of chatgpt on nlp tasks. arXiv preprint arXiv:2306.09719, 2023.
13. Ren X, Wei W, Xia L, et al. Representation Learning with Large Language Models for Recommendation. Published online February 25, 2024. doi:10.1145/3589334.3645458
14. Wei J, Tay Y, Bommasani R, et al. Emergent Abilities of Large Language Models. Published online October 26, 2022. Accessed August 30, 2023. http://arxiv.org/abs/2206.07682
15. Jiang Z, Xu F, Gao L, et al. Active Retrieval Augmented Generation. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics; 2023:7969-7992. doi:10.18653/v1/2023.emnlp-main.495
16. Wang Z, Teo S, Ouyang J, Xu Y, Shi W. M-RAG: Reinforcing Large Language Model Performance through Retrieval-Augmented Generation with Multiple Partitions. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics; 2024:1966-1978. doi:10.18653/v1/2024.acl-long.108
17. Zhang Z, Han X, Liu Z, et al. ERNIE: Enhanced Language Representation with Informative Entities. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 1441-1451.
18. Peters ME, Neumann M, Logan IV RL, et al. Knowledge Enhanced Contextual Word Representations. Published online October 30, 2019. Accessed August 24, 2023.

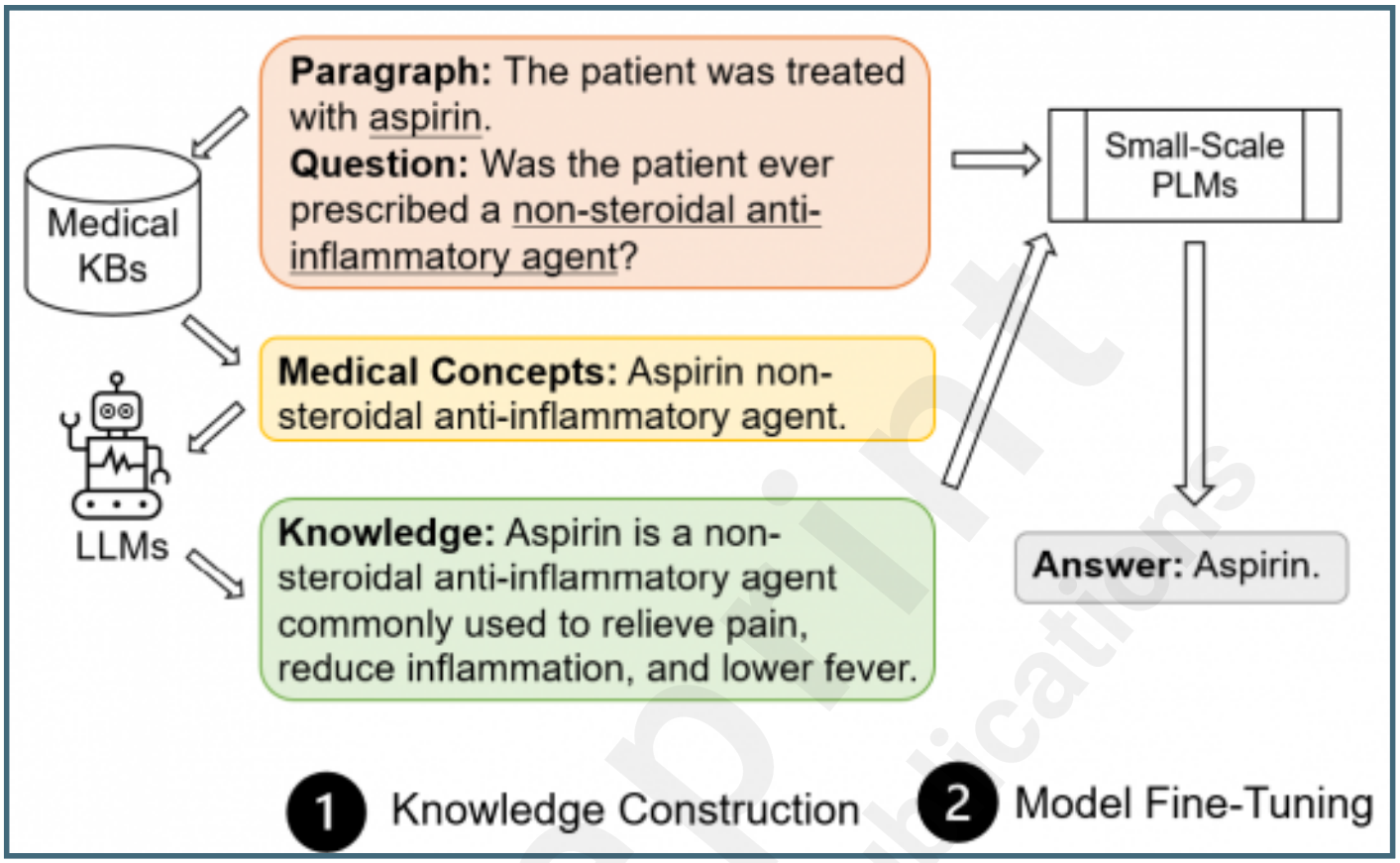
http://arxiv.org/abs/1909.04164
19. Xiong W, Du J, Wang WY, Stoyanov V. Pretrained Encyclopedia: Weakly Supervised Knowledge-Pretrained Language Model. Published online December 19, 2019. Accessed August 24, 2023. http://arxiv.org/abs/1912.09637
20. Wang R, Tang D, Duan N, et al. K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters. Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. 2021: 1405-1418.
21. Sun T, Shao Y, Qiu X, et al. CoLAKE: Contextualized Language and Knowledge Embedding. In: Proceedings of the 28th International Conference on Computational Linguistics. International Committee on Computational Linguistics; 2020:3660-3670. doi:10.18653/v1/2020.coling-main.327
22. Wang X, Gao T, Zhu Z, et al. KEPLER: A unified model for knowledge embedding and pre-trained language representation. Transactions of the Association for Computational Linguistics, 2021, 9: 176-194.
23. Cui L, Wu Y, Liu S, Zhang Y. Knowledge Enhanced Fine-Tuning for Better Handling Unseen Entities in Dialogue Generation. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics; 2021:2328-2337. doi:10.18653/v1/2021.emnlp-main.179
24. He Y, Zhu Z, Zhang Y, Chen Q, Caverlee J. Infusing Disease Knowledge into BERT for Health Question Answering, Medical Inference and Disease Name Recognition. Published online October 7, 2020. Accessed June 2, 2024. http://arxiv.org/abs/2010.03746
25. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Wren J, ed. Bioinformatics. 2020;36(4):1234-1240. doi:10.1093/bioinformatics/btz682
26. Huang K, Altosaar J, Ranganath R. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. Published online November 29, 2020. Accessed October 25, 2024. http://arxiv.org/abs/1904.05342
27. Remy F, Demuynck K, Demeester T. BioLORD: Learning Ontological Representations from Definitions (for Biomedical Concepts and their Textual Descriptions). Published online October 21, 2022. Accessed June 2, 2024. http://arxiv.org/abs/2210.11892
28. Wang X, Yang Q, Qiu Y, et al. KnowledGPT: Enhancing Large Language Models with Retrieval and Storage Access on Knowledge Bases. Published online August 17, 2023. doi:10.48550/arXiv.2308.11761
29. Wang S, Xu Y, Fang Y, et al. Training Data is More Valuable than You Think: A Simple and Effective Method by Retrieving from Training Data. Published online March 16, 2022. doi:10.48550/arXiv.2203.08773
30. Cheng X, Luo D, Chen X, Liu L, Zhao D, Yan R. Lift Yourself Up: Retrieval-augmented Text Generation with Self-Memory. Adv Neural Inf Process Syst. 2023;36:43780-43799.
31. Lin XV, Chen X, Chen M, et al. RA-DIT: Retrieval-Augmented Dual Instruction Tuning. Published online May 6, 2024. doi:10.48550/arXiv.2310.01352
32. Cheng D, Huang S, Bi J, et al. UPRISE: Universal Prompt Retrieval for Improving Zero-Shot Evaluation. Published online December 16, 2023. doi:10.48550/arXiv.2303.08518
33. Dai Z, Zhao VY, Ma J, et al. Promptagator: Few-shot Dense Retrieval From 8 Examples. Published online September 23, 2022. doi:10.48550/arXiv.2209.11755
34. Asai A, Wu Z, Wang Y, Sil A, Hajishirzi H. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. Published online October 17, 2023. doi:10.48550/arXiv.2310.11511
35. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004;32(90001):267D - 270. doi:10.1093/nar/gkh061

36. Neumann M, King D, Beltagy I, Ammar W. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In: Proceedings of the 18th BioNLP Workshop and Shared Task. Association for Computational Linguistics; 2019:319-327. doi:10.18653/v1/W19-5034
37. Touvron H, Martin L, Stone K, et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. Published online July 19, 2023. Accessed August 29, 2023. http://arxiv.org/abs/2307.09288
38. Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res. 2020;21(140):1-67.
39. Pampari A, Raghavan P, Liang J, Peng J. emrQA: A Large Corpus for Question Answering on Electronic Medical Records. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics; 2018:2357-2368. doi:10.18653/v1/D18-1258
40. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Published online May 24, 2019. Accessed September 8, 2023. http://arxiv.org/abs/1810.04805

Supplementary Files

Figures

An Example of RAG-Based Model Fine-Tuning. The underlined words in the question and paragraph represent extracted medical concepts.



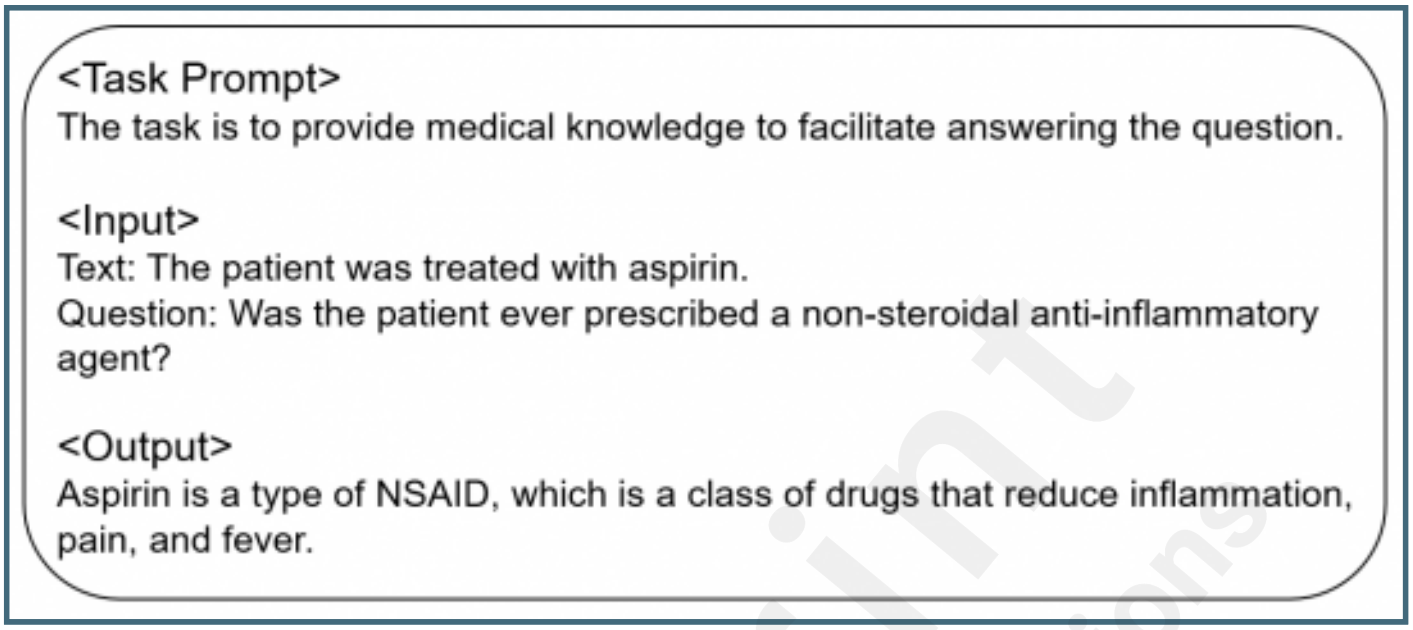
The Prompt Template of RAG-Based Knowledge Construction.

<Task Prompt>
The task is to generate a natural sentence based on the given text.

<Input>
Medical Concepts: Aspirin non-steroidal anti-inflammatory agent.

<Output>
Knowledge: Aspirin is a non-steroidal anti-inflammatory agent commonly used to relieve pain, reduce inflammation, and lower fever.

The Prompt Template of LLM-Based Knowledge Construction.



The Prompt Template of QA Task. The underlined sentence represents the inserted knowledge in knowledge enhancement scenarios.

<Task Prompt>
The task is to answer the question based on the paragraph.

<Input>
Paragraph: The patient was treated with aspirin. Aspirin is a non-steroidal anti-inflammatory agent commonly used to relieve pain, reduce inflammation, and lower fever.
Question: Was the patient ever prescribed a non-steroidal anti-inflammatory agent?

<Output>
Answer: Aspirin.