KG-Rank: Enhancing Large Language Models for Medical QA with Knowledge Graphs and Ranking Techniques

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have significantly advanced healthcare innovation on generation capabilities. However, their application in real clinical settings is challenging due to potential deviations from medical facts and inherent biases. In this work, we develop an augmented LLM framework, KG-Rank, which leverages a medical knowledge graph (KG) with ranking and re-ranking techniques, aiming to improve free-text question-answering (QA) in the medical domain. Specifically, upon receiving a question, we initially retrieve triplets from a medical KG to gather factual information. Subsequently, we innovatively apply ranking methods to refine the ordering of these triplets, aiming to yield more precise answers. To the best of our knowledge, KG-Rank is the first application of ranking models combined with KG in medical QA specifically for generating long answers. Evaluation of four selected medical QA datasets shows that KG-Rank achieves an improvement of over 18% in the ROUGE-L score. Moreover, we extend KG-Rank to open domains, where it realizes a 14% improvement in ROUGE-L, showing the effectiveness and potential of KG-Rank.

1 Introduction

Large language models (LLMs) such as GPT-4 (OpenAI, 2023) and LLaMa2 (Touvron et al., 2023) have demonstrated powerful text generation capabilities. While they hold vast promise in various domains, including healthcare (Yang et al., 2023b; Li et al., 2022; Ke et al., 2024), their limited training on medical data raises concerns about generating medically inaccurate content and perpetuating biases based on ethnicity, gender, and socioeconomic status (Xie et al., 2023; Yang et al., 2023a).

To address these challenges without additional computation cost, previous research like Almanac (Hiesinger et al., 2023) and ChatENT (Long et al., 2023) incorporate external medical knowledge bases and show that can enhance LLM accuracy and reliability. However, these methods risk introducing irrelevant or unreliable information, compromising the LLM's effectiveness, and raising issues of credibility, data consistency, privacy, security, and legality. While previous studies have emphasized the advantages of utilizing external knowledge bases, they neglect the question: *How to better integrate knowledge base contents?*

In this work, we propose **KG-Rank**, a framework that integrates a structured medical knowledge graph, into existing pre-trained LLMs to achieve more accurate medical question-answering (QA). We first retrieve one-hop relations of related medical entities from a reliable medical KG, the Unified Medical Language System (UMLS) database (Bodenreider, 2004). Then, to keep accurate and helpful information from the KG, we propose to apply ranking and re-ranking methods to optimize the triplets ordering. Specifically, we introduce three ranking methods to improve LLM integration by filtering irrelevant data, highlighting key information, ensuring diversity, and boosting the system's factuality. These techniques also streamline the process by reducing the number of triplets required for LLM inference. Additionally, we develop re-ranking methods to reassess and emphasize the most relevant triplets, further refining the QA performance.

To summarize, our contributions are: (1) We propose KG-Rank, a KG-enhanced LLM framework for medical QA tasks. To the best of our knowledge, this is the first application of KG and ranking-enhanced LLMs to medical QA with long answers. (2) We incorporate ranking techniques to improve factuality and eliminate noise and redundancy in the KG-retrieval stage. (3) We validate our model's effectiveness on both medical and open domain QA tasks. Code can be found in `https://anonymous.4open.science/r/KGQA-270F`.

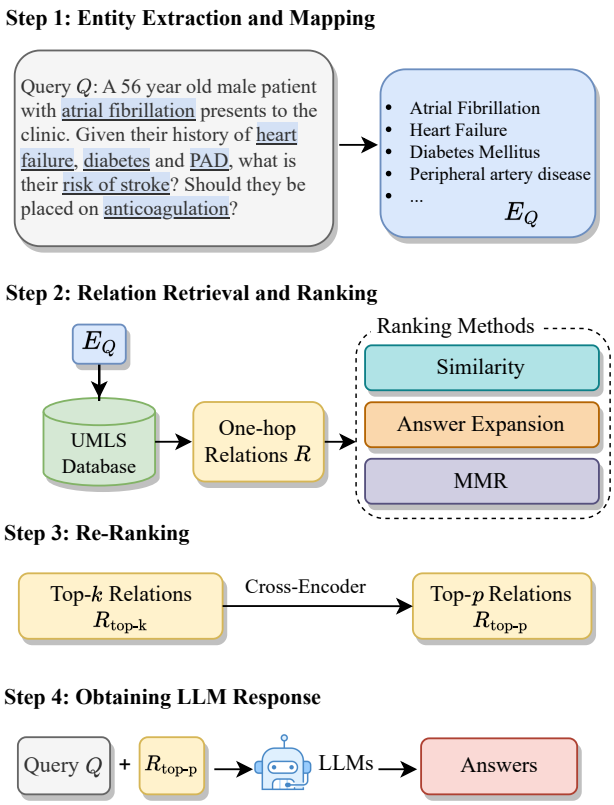

Figure 1: An illustration of KG-Rank Framework.

2 Methodology

Shown in Fig. 1, We introduce the KG-Rank (Knowledge Graph-Rank) framework for QA tasks.

2.1 External Knowledge Graph

We define the external KG as $G = (V, E)$, where V represents the set of medical entities and E represents the set of structural relations. For medical QA, we choose UMLS (Bodenreider, 2004) as the primary medical KG. UMLS is a repository of health and biomedical vocabularies. The core component of UMLS, Metathesaurus, contains over 3.8 million concepts and more than 78 million relations, and supports 25 languages, providing extensive medical knowledge coverage to enhance LLMs.

2.2 Entity Extraction and Mapping

In the first step, we extract key entities and find mappings from the external KG. Specifically, for the given question Q, we apply a Medical NER Prompt P_{MedNER} to identify related medical entities E. We map each entity $e_i \in E$ to the corresponding entity in the knowledge graph G.

2.3 Relation Retrieval and Ranking

After identifying the corresponding entities E_Q, we retrieve their one-hop relations from the KG (denoted as *UMLS Database*):

$$E_Q = \{e'_i \in V \mid \exists e_i \in E, e_i \mapsto e'_i\}.$$

Within UMLS, there exists extensive relational information, where one entity may be associated with thousands of one-hop relations. Consequently, to facilitate the extraction of the most relevant, we propose ranking methods. We encode the question Q and each triplet (e'_i, r, e'_j) into $\mathbf{q}, \mathbf{r}_{ij}$ through UmlsBERT (Michalopoulos et al., 2021). Then, we explore three techniques for ranking the triplets:

Similarity Ranking We compute the similarity score between the question embedding $\mathbf{q}$ and each relation embedding $\mathbf{r}_{ij}$.

Answer Expansion Ranking We first utilize LLMs to generate a hallucinatory answer A for the question Q , and then we encode the concatenation of $[Q, A]$ to obtain text embedding $\mathbf{t}$. Subsequently, we utilize the expanded question embedding $\mathbf{t}$ to search for the most similar triplets in vector space.

MRR Ranking This method is inspired by an information extraction method Maximal Marginal Relevance (MMR) (Carbonell and Goldstein-Stewart, 1998). Initially, we identify the triplet with the highest similarity score to the question Q. For the remaining triplets, we dynamically adjust their similarity scores based on the ones that have already been selected. In this way, we could consider both relevancy and redundancy:

$$w = w_{base} + \delta \cdot n,$$

$$\text{score}_{ij} = \text{sim}(\mathbf{q}, \mathbf{r}_{ij}) - w \cdot \overline{\text{sim}}(\mathbf{r}_{ij}, \mathbf{r}_{sel}).$$

Where, w is an adjustable weight, with a base weight and δ as the incremental weight factor per selected triplet, n is the count of triplets that have been selected.

Re-ranking After the ranking stage, we obtain an ordering of the triplets. We then employ a medical cross-encoder model to re-rank them, ensuring that the most relevant triples are chosen. The re-ranked top-p triplets, combined with the task prompt, are input into LLMs for answer generation.

3 Experiments

3.1 Datasets

We conduct experiments on four selected medical QA datasets, in which the answers are free-text.

Dataset	Method	GPT-4				LLaMA2-13b			
		ROUGE-L	BERTScore	MoverScore	BLEURT	ROUGE-L	BERTScore	MoverScore	BLEURT
LiveQA	ZS	18.89	82.50	54.02	39.84	17.73	81.93	53.37	40.45
	Sim	19.35	**83.01**	54.08	40.47	18.52	82.78	**53.79**	**40.59**
	AE	19.24	82.95	54.04	40.15	18.45	82.60	53.70	39.80
	MMR	19.32	82.91	54.03	**40.55**	18.25	82.70	53.67	40.22
	RR	**19.44**	82.94	**54.11**	40.50	**18.83**	**82.79**	53.72	39.59
ExpertQA-Bio	ZS	23.00	84.50	56.15	44.53	23.26	84.38	55.58	44.65
	Sim	25.90	85.72	56.73	45.10	24.96	84.91	55.83	44.35
	AE	26.78	85.77	56.79	45.18	24.84	84.97	55.72	43.55
	MMR	26.54	85.76	56.77	44.93	25.40	85.08	55.98	44.04
	RR	**27.20**	**85.83**	**57.11**	**45.91**	**25.79**	**85.18**	**56.17**	**45.20**
ExpertQA-Med	ZS	25.45	85.11	56.50	45.98	24.86	84.89	55.74	46.32
	Sim	27.61	86.10	57.13	46.47	26.40	85.50	56.23	46.15
	AE	27.98	86.12	57.25	46.80	26.15	85.36	56.17	46.02
	MMR	27.78	86.22	57.28	46.84	26.42	85.57	56.24	46.41
	RR	**28.08**	**86.30**	**57.32**	**47.00**	**27.49**	**85.80**	**56.58**	**46.47**
MedicationQA	ZS	14.41	82.55	52.62	37.41	13.30	81.81	51.96	38.30
	Sim	16.05	83.56	53.23	37.60	14.60	82.73	52.47	38.38
	AE	16.13	83.46	53.23	37.87	14.19	82.50	52.33	37.90
	MMR	15.89	83.48	53.22	37.73	14.56	82.69	52.44	38.31
	RR	**16.19**	**83.59**	**53.30**	**37.91**	**14.71**	**82.79**	**52.59**	**38.42**

Table 1: Automatic evaluation scores: we compare ROUGE-L, BERTScore, MoverScore, BLEURT on different settings. The superior scores among the same models are highlighted in **bold**.

LiveQA (Abacha et al., 2017) consists of health questions submitted by consumers to NLM. It includes a training set with 634 QA pairs and a test set comprising 104 QA pairs, which is used for evaluation. ExpertQA (Malaviya et al., 2023) is a high-quality long-form QA dataset with 2177 questions spanning 32 fields, along with answers verified by domain experts. Among them, 504 medical questions (Med) and 96 biology (Bio) questions were used for evaluation. MedicationQA (Abacha et al., 2019) includes 690 drug-related consumer questions along with information retrieved from reliable websites and scientific papers. We evaluate the generated answers using ROUGE (Lin, 2004), BERTScore (Zhang et al., 2019), MoverScore (Zhao et al., 2019) and BLEURT (Sellam et al., 2020).

3.2 Results

As shown in Tab. 1, we evaluate GPT-4 and LLaMa2-13b across the following settings: zero-shot (ZS) , and three proposed ranking techniques: Similarity Ranking (Sim), Answer Expansion (AE), and Maximal Marginal Relevance Ranking (MMR). Also with the Re-ranking (RR), which is on top of the Similarity Ranking.

The results show that incorporating the knowledge graph and ranking methods notably enhances performance in almost all benchmarks and evaluation metrics in the zero-shot setting, demonstrating the effectiveness of KG-Rank. Significantly, the RR method excels in the ExpertQA-Bio, ExpertQA-Med, and Medication QA datasets, particularly evident in the over 18% increase in the ROUGE-L score for ExpertQA-Bio. While KG-Rank still shows effectiveness on LiveQA, the RR method is not improving steadily among other ranking methods. This inconsistency may arise since the answers in LiveQA are generated via automatic extraction methods, leading to issues with semantic coherence and disorganized formats. Moreover, the performance of the three ranking methodologies exhibited variability across various datasets, indicating their unique strengths and limitations in differing contexts.

Dataset	GPT-4			
	ROUGE-L	BERTScore	MoverScore	BLEURT
Cohere				
LiveQA	18.72	82.94	54.08	40.07
ExpertQA-Bio	26.08	85.81	56.93	45.70
ExpertQA-Med	27.59	86.08	57.14	46.54
MedicationQA	16.14	83.46	53.25	37.82
MedCPT				
LiveQA	**19.44**	**82.95**	**54.11**	**40.50**
ExpertQA-Bio	**27.20**	**85.83**	**57.11**	**45.91**
ExpertQA-Med	**28.08**	**86.30**	**57.32**	**46.84**
MedicationQA	**16.19**	**83.59**	**53.30**	**37.91**

Table 2: The performance of Cohere re-rank model and MedCPT in the re-ranking stage.

In assessing model performance, GPT-4 consistently surpasses LLaMa2-13b in both zero-shot and various ranking settings. Additionally, we evaluate the zero-shot performance of a medical LLM (Xu et al., 2023) on these datasets.

4 Ablation Study and Analysis

Medical LLMs. To further investigate the capabilities of medical LLMs, we compare the zero-shot performance of LLaMa2-7b and baize-healthcare

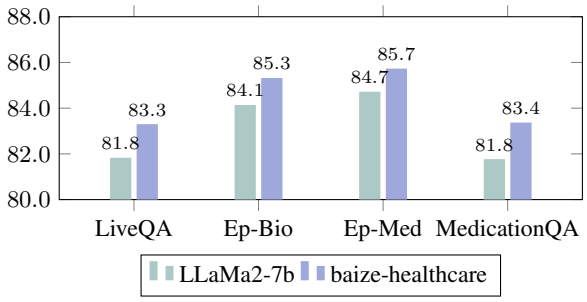

Figure 2: BERTScore Comparison: Zero-shot with LLaMa2-7b and Baize-Healthcare. Ep stands for ExpertQA.

Question: What would be the ideal diet recommendations for a 53 year old male with acute renal failure and hepatic failure?

LLaMa2-13b: ...3. Protein intake: Provide adequate protein to support liver function and wound healing. The recommended daily intake of protein for a 53-year-old male is approximately 1.6-2.2 grams per kilogram of body weight...

LLaMa2-13b w KG-Rank:...1. Low protein intake: Protein intake should be limited to about 0.8-1.0 gram per kilogram of body weight per day, as excessive protein intake can exacerbate renal failure and liver disease...

Figure 3: A case study from ExpertQA-Med: results from LLaMa2-13b and with KG-Rank.

using BERTScore. Baize-healthcare, which is fine-tuned on LLaMa-7b using medical data, consistently outperforms LLaMa2-7b across all four datasets, as shown in Fig. 2. More results are in Appendix B.1.

Re-ranking Models. We employ GPT-4 for the final answer generation and compare two re-ranking models: the MedCPT cross-encoder model, trained on the extensive PubMed dataset, and the Cohere (`https://cohere.com`) re-ranking model, designed for broader domain applications. As shown in Tab. 2, we find that MedCPT steadily outperforms the Cohere re-rank model on all datasets, highlighting the importance of specialized models in the medical field. Additional evaluations are provided in Appendix B.2.

Case Study. To further analyze the output under the KG-Rank framework, a case study is presented in Fig. 3. When asked about ideal diet recommendations for a 53-year-old male with acute renal failure and hepatic failure, both provide guidelines regarding protein intake. However, the original recommendation emphasizes ensuring adequate protein consumption (*1.6-2.2 grams per kilogram*), whereas the answer generated under the KG-Rank framework advises controlling protein intake (*limited to about 0.8-1 gram per kilogram*). The difference is critical for patients with acute renal and hepatic failure, where an inappropriate protein dosage, such as the higher range of 1.6-2.2 grams per kilogram, could worsen the strain on already compromised kidneys and liver, potentially leading to escalated health issues. This case shows that KG-Rank is more factually correct in the generated answer.

Setting	ROUGE-L	BERTScore	MoverScore
ExpertQA-Law			
Base	26.33	85.03	48.57
KG-Rank	**29.93**	**86.25**	**48.63**
ExpertQA-Business			
Base	21.78	84.46	48.92
KG-Rank	**24.20**	**85.42**	**49.10**
ExpertQA-Music			
Base	23.84	85.21	45.73
KG-Rank	**27.31**	**86.23**	**46.55**
ExpertQA-History			
Base	25.65	85.55	45.82
KG-Rank	**27.75**	**86.21**	**47.08**

Table 3: Base and KG-Rank performance in the open domain.

KG-Rank in Open Domain Additionally, to demonstrate the effectiveness of our KG-Rank, we extend it to the open domain by replacing UMLS with Wikipedia through the DBpedia API (`https://www.dbpedia.org/`). We experiment on Mintaka (Sen et al., 2022), and randomly select 1,000 pairs from the test set for evaluation. We improve the accuracy from 60.40% to 61.90%. In Tab. 3, we also conduct experiments in the four fields of Law, Business, Music, and History within ExpertQA, using GPT-4 as the base model. Results show that KG-Rank outperforms baseline across all benchmarks. Building on these findings, the efficacy of our framework is not limited to the medical domain, and has the potential to be applied in a variety of areas. For more case studies, please refer to the Appendix C.

5 Conclusion

In this work, we proposed KG-Rank, an enhanced LLM framework combining medical knowledge graphs with ranking techniques to boost free-text medical QA. KG-Rank, as far as we know, is the inaugural integration of ranking models with KG for long-answer medical QA. It demonstrates over 18% improvement in ROUGE-L across four medical QA datasets. Its application to open domains yields a 14% ROUGE-L score enhancement, underscoring KG-Rank's effectiveness and versatility.

Limitations

In this research, we propose an LLM framework augmented by UMLS to improve the quality of the content generated. However, there are some limitations, which we will address in the next phase. Firstly, we plan to incorporate physician evaluations to validate the factual accuracy of KG-Rank's answers. Secondly, we aim to assess the performance of more medical-specific base models on medical QA tasks. Lastly, while the ranking method may increase computational time, we recognize the need to optimize its efficiency.

Ethical Considerations

This research utilized public medical datasets solely for academic purposes, not for practical application. We employed GPT-4, LLaMa2-13b, LLaMa2-7b, baize-healthcare for text generation, ensuring that no harmful content was produced. Both the benchmark datasets and the model outputs are free of any individual privacy data.

References

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

Asma Ben Abacha, Yassine Mrabet, Mark Sharp, Travis R Goodwin, Sonya E Shooshan, and Dina Demner-Fushman. 2019. Bridging the gap between consumers' medication questions and trusted answers. In *MedInfo*, pages 25–29.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Jaime G. Carbonell and Jade Goldstein-Stewart. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

William Hiesinger, Cyril Zakka, Akash Chaurasia, Rohan Shad, Alex Dalal, Jennifer Kim, Michael Moor, Kevin Alexander, Euan Ashley, Jack Boyd, et al. 2023. Almanac: Retrieval-augmented language models for clinical medicine.

Yu He Ke, Rui Yang, Sui An Lie, Taylor Xin Yi Lim, Hairil Rizal Abdullah, Daniel Shu Wei Ting, and Nan Liu. 2024. Enhancing diagnostic accuracy through multi-agent conversations: Using large language models to mitigate cognitive bias.

Irene Li, Jessica Pan, Jeremy Goldwasser, Neha Verma, Wai Pan Wong, Muhammed Yavuz Nuzumlalı, Benjamin Rosand, Yixin Li, Matthew Zhang, David Chang, et al. 2022. Neural natural language processing for unstructured data in electronic health records: A review. *Computer Science Review*, 46:100511.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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

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

George Michalopoulos, Yuanxin Wang, Hussam Kaka, Helen Chen, and Alexander Wong. 2021. Umlsbert: Clinical domain knowledge augmentation of contextual embeddings using the unified medical language system metathesaurus.

OpenAI. 2023. Gpt-4 technical report.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Priyanka Sen, Alham Fikri Aji, and Amir Saffari. 2022. Mintaka: A complex, natural, and multilingual dataset for end-to-end question answering.

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

Qianqian Xie, Edward J Schenck, He S Yang, Yong Chen, Yifan Peng, and Fei Wang. 2023. Faithful ai in medicine: A systematic review with large language models and beyond. *Medrxiv: the Preprint Server for Health Sciences*.

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *arXiv preprint arXiv:2304.01196*.

Rui Yang, Ting Fang Tan, Wei Lu, Arun James Thirunavukarasu, Daniel Shu Wei Ting, and Nan Liu. 2023a. Large language models in health care: Development, applications, and challenges. *Health Care Science*.

Rui Yang, Qingcheng Zeng, Keen You, Yujie Qiao, Lucas Huang, Chia-Chun Hsieh, Benjamin Rosand, Jeremy Goldwasser, Amisha D Dave, Tiarnan D. L. Keenan, Emily Y Chew, Dragomir Radev, Zhiyong

Lu, Hua Xu, Qingyu Chen, and Irene Li. 2023b. Ascle: A python natural language processing toolkit for medical text generation.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. *arXiv preprint arXiv:1909.02622*.

A Prompt Templates

In this section, we present the detailed prompt templates employed as inputs for LLMs at each phase of the KG-Rank process.

A.1 Medical NER Prompt

Fig. 4 illustrates the Medical NER prompt template that is specifically designed for extracting medical terminologies from a given question.

Question: {question}

You are interacting with a knowledge graph that contains definitions and relational information of medical terminologies. To provide a precise and relevant answer to this question, you are expected to:

1. Understand the Question Thoroughly: Analyze the question deeply to identify which specific medical terminologies and their interrelations, as extracted from the knowledge graph, are crucial for formulating an accurate response.
2. Extract Key Terminologies: Return the 3-5 most relevant medical terminologies based on their significance to the question.
3. Format the Output: Return in a structured JSON format with the key as "medical terminologies".

For example:
{"medical terminologies": ["term1", "term2", ...]}

Figure 4: Prompt used to extract medical terminologies.

A.2 Answer Expansion Prompt

Figure 5 illustrates the prompt template designed for our proposed answer expansion ranking strategy, as shown in step 2 of Fig. 1 and as described in Section 2.3.

Question: {question}

Provide an example answer to the given question.

Your answer is derived from a biomedical knowledge graph.

This knowledge graph encompasses a wide range of medical terminologies and elucidates the complex interconnections between these terms, supporting an in-depth and accurate response to the question.

Figure 5: Prompt for answer expansion ranking method.

A.3 KG-enhanced Prompt

Fig. 6 shows the prompt template to obtain final answers from LLMs, corresponding to step 4 in Fig. 1.

Answer the question in conjunction with the following content.

Context:
{context}

Patient:
{input}

Physician:

Figure 6: Prompt for obtaining KG-enhanced LLM answers.

A.4 KG-enhanced Prompt for Mintaka task

Fig. 7 presents the prompt for obtaining KG-enhanced LLM answers, specially designed for the Mintaka dataset, as it's a multi-choice QA dataset.

Here are some examples for output format:

Question: What is the seventh tallest mountain in North America?
Example Output: Mount Lucania

Question: What year was the first book of the A Song of Ice and Fire series published?
Example Output: 1996

Question: How old was Taylor Swift when she won her first Grammy?
Example Output: 20

Question: Has there ever been a Christian U.S. senator?
Example Output: Yes

Context:
{context}

Question:
{input}

Answer:

Figure 7: Prompt for obtaining KG-enhanced LLM answers, with special design for Mintaka dataset.

B Detailed Evaluation Results

B.1 Zero-shot Performance of Different LLMs

In this section, we evaluate the zero-shot Medical QA performance of widely-used LLMs. As shown in Tab. 4, the results demonstrate that GPT-4 performing better than the other models, which suggests it's generally more accurate. The baize-healthcare model scores the highest in BERTScore on all datasets, indicating its superior text comprehension abilities. On the other hand, LLaMa2-13b achieves the top scores in BLEURT, implying its effectiveness on producing answers that read well and match the facts closely. This comparison highlights what each model does best when answering medical questions without any external reference.

Dataset	Evaluation Metrics					
	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	MoverScore	BLEURT
LLaMa2-7b						
LiveQA	18.87	3.60	17.44	81.83	53.28	39.43
ExpertQA-Bio	24.19	6.96	22.15	84.14	55.18	43.81
ExpertQA-Med	26.24	8.11	23.86	84.72	55.51	45.75
MedicationQA	14.19	2.60	13.12	81.77	51.94	37.32
baize-healthcare						
LiveQA	17.92	2.73	16.10	**83.30**	53.41	31.30
ExpertQA-Bio	23.45	6.52	21.31	**85.32**	54.95	33.80
ExpertQA-Med	24.95	7.21	22.41	**85.73**	55.12	34.52
MedicationQA	15.05	2.48	13.59	**83.37**	52.41	31.39
LLaMa2-13b						
LiveQA	19.15	3.60	17.73	81.93	53.37	**40.45**
ExpertQA-Bio	**25.33**	**7.92**	**23.26**	84.38	55.58	**44.65**
ExpertQA-Med	27.41	8.86	24.86	84.89	55.74	**46.32**
MedicationQA	14.42	2.62	13.30	81.81	51.96	**38.30**
GPT-4						
LiveQA	**20.54**	**4.65**	**18.89**	82.50	**54.02**	39.84
ExpertQA-Bio	25.06	7.84	23.00	84.50	**56.15**	44.53
ExpertQA-Med	**27.78**	**9.49**	**25.45**	85.11	**56.50**	45.98
MedicationQA	**15.52**	**3.51**	**14.41**	82.55	**52.62**	37.41

Table 4: Automatic evaluation scores: we compare ROUGE-1, ROUGE-2, ROUGE-L, BERTScore, MoverScore, BLEURT on the zero-shot setting for different LLMs with medical QA tasks. The best scores among the same models are highlighted in **bold**.

B.2 Performance of Different Re-rank Models

Dataset	GPT-4					
	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	MoverScore	BLEURT
Cohere						
LiveQA	21.08	4.13	18.72	82.94	54.08	40.07
ExpertQA-Bio	29.07	9.35	26.08	85.81	56.93	45.70
ExpertQA-Med	30.84	10.62	27.59	86.08	57.14	46.54
MedicationQA	17.76	3.65	16.14	83.46	53.25	37.82
MedCPT						
LiveQA	**21.70**	**4.33**	**19.44**	**82.95**	**54.11**	**40.50**
ExpertQA-Bio	**30.05**	**10.51**	**27.20**	**85.83**	**57.11**	**45.91**
ExpertQA-Med	**31.34**	**10.96**	**28.08**	**86.30**	**57.32**	**46.84**
MedicationQA	**17.94**	**3.72**	**16.19**	**83.59**	**53.30**	**37.91**

Table 5: Automatic evaluation scores: we compare the performance of different re-rank models on ROUGE-1, ROUGE-2, ROUGE-L, BERTScore, MoverScore, BLEURT. The best scores among the same models are highlighted in **bold**.

C More Case Studies

We put another case study from the ExpertQA-Med dataset, where in regards to the prognosis survival rates of breast cancer cases, the answer generated by KG-Rank is more factually accurate in terms of medical evidence, as shown in Fig. 8. Moreover, Fig. 9 shows a case study on the open-domain QA tasks from the Mintaka dataset, comparing the performance of the vanilla GPT-4 model against the KG-Rank-enhanced GPT-4 model. The case study involves a question: "How many of the Godfather movies was Robert De Niro in?" While GPT-4 responded with "2", our proposed KG-Rank-enhanced GPT-4 provided the correct answer "1", which matches the ground truth. We also show the evidence retrieved from DBPedia. This case study shows that by incorporating KG-Rank, the model is able to leverage the relevant information effectively to derive the correct answer, whereas the vanilla GPT-4 did not. This demonstrates the efficacy of KG-Rank in improving the accuracy of answers in LLMs when dealing with general domain factual questions.

D Experimental Setup

In our experimental setup, we employ `UmlsBERT`[1], `baize-healthcare`[2], `llama-2-7b-chat-hf`[3], `llama-2-13b-chat-hf`[4], `MedCPT`[5] from Hugging Face. For GPT-4, we use the OpenAI API with a zero-temperature setting. For the Cohere re-rank model, we employ it through its API. All experiments are conducted on a cluster equipped with 4 NVIDIA A100 GPUs. The prediction for each sample takes about a few seconds. Based on the size of each dataset, it may take up to hours to finish the evaluation.

[1] GanjinZero/UMLSBert_ENG
[2] https://huggingface.co/project-baize/baize-healthcare-lora-7B
[3] https://huggingface.co/meta-llama
[4] https://huggingface.co/meta-llama
[5] https://huggingface.co/ncbi/MedCPT-Cross-Encoder

Question:
What is the prognosis of a case of triple negative ductal breast carcinoma?

LLaMa2-13b:

The prognosis for triple negative ductal breast cancer (TNBC) varies depending on several factors, such as the stage of the disease, the presence of certain biomarkers, and the patient's overall health. However, in general, the prognosis for TNBC is considered to be less favorable compared to other types of breast cancer.

Here are some general prognostic factors for TNBC:
1. Stage: The stage of the cancer is a major determinant of prognosis. TNBC that is confined to the breast (stage I or II) has a better prognosis than more advanced stages (III or IV).
2. Lymph node involvement: If the cancer has spread to the lymph nodes under the arm (axillary lymph nodes), it can indicate a poorer prognosis.
3. Biomarkers: The presence of certain biomarkers, such as high levels of Ki-67 (a protein that indicates cell division) or low levels of estrogen receptors (ER) and progesterone receptors (PR), can indicate a poorer prognosis.
4. Genetic mutations: The presence of specific genetic mutations, such as those in the BRCA1 or BRCA2 genes, can increase the risk of recurrence and reduce the prognosis.
5. Patient characteristics: Factors such as age, overall health, and the presence of other medical conditions can also affect the prognosis for TNBC patients.

Despite these factors, many women with TNBC can achieve long-term survival and even cure. According to the American Cancer Society, the 5-year survival rate for TNBC is about 70% for women who are diagnosed at an early stage (stage I or II). For women diagnosed at a later stage (stage III or IV), the 5-year survival rate is about 40%.

It's important to note that these statistics are based on averages and that each individual case is unique. Many factors can influence the prognosis for TNBC, and it's best to consult with a doctor to get a more accurate assessment of your specific situation.

LLaMa2-13b w KG-Rank:

The prognosis of a case of triple negative ductal breast carcinoma can be challenging to determine due to the lack of specific targeted therapies and the limited availability of clinical trials for this specific subtype of breast cancer.

However, some studies suggest that triple negative breast cancer patients who are diagnosed at an early stage (stage I or II) have a relatively good prognosis, with a 5-year survival rate of around 80-90%. However, for patients with more advanced stages (stage III or IV), the prognosis is generally poorer, with a 5-year survival rate of around 50-60%. This is because advanced stages of triple negative breast cancer are often associated with a higher risk of recurrence and metastasis.

It's important to note that the prognosis for triple negative breast cancer can vary depending on several factors, such as the size of the tumor, the presence of lymph node involvement, and the patient's overall health status. Additionally, the prognosis can also be influenced by the specific molecular biomarkers present in the tumor, such as the ERBB2 gene, which has been shown to be associated with a worse prognosis in some studies.

Overall, while the prognosis for triple negative ductal breast carcinoma can be challenging to predict, it is important for patients to receive prompt and appropriate treatment, including surgery, chemotherapy, and targeted therapies where available, to improve their chances of a successful outcome.

Figure 8: A case study from ExpertQA-Med: we show results from vanilla LLaMa2-13b and KG-Rank-enhanced LLaMa2-13b.

Question:
How many of the Godfather movies was Robert De Niro in?

GPT-4 Answer: 2
KG-Rank: 1
Ground_truth: 1

Retrieved from Wiki: Sian unuan Akademian Premion Robert de Niro gajnis en la jaro 1974, pro la plej bona epizoda rolo en la filmo The Godfather, Part II, direktita de Francis Ford Coppola.

Figure 9: A case study from Mintaka: we show results from vanilla GPT-4 and KG-Rank-enhanced GPT-4.