

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

Rui Yang^{1*}, Haoran Liu², Edison Marrese-Taylor³, Qingcheng Zeng⁴, Yu He Ke⁵, Wanxin Li⁶, Lechao Cheng⁷, Qingyu Chen^{8,9}, James Caverlee², Yutaka Matsuo³, Irene Li^{3*}

¹Duke-NUS Medical School, ²Texas A&M University, ³The University of Tokyo,

⁴Northwestern University, ⁵Singapore General Hospital, ⁶Zhejiang University,

⁷Zhejiang Lab, ⁸Yale University, ⁹National Institutes of Health

yang.rui@duke-nus.edu.sg, ireneli@ds.itc.u-tokyo.ac.jp

Abstract

Large language models (LLMs) have demonstrated impressive generative capabilities with the potential to innovate in medicine. However, the application of LLMs in real clinical settings remains challenging due to the lack of factual consistency in the generated content. In this work, we develop an augmented LLM framework, KG-Rank, which leverages a medical knowledge graph (KG) along with ranking and re-ranking techniques, to improve the factuality of long-form question answering (QA) in the medical domain. Specifically, when receiving a question, KG-Rank automatically identifies medical entities within the question and retrieves the related triples from the medical KG to gather factual information. Subsequently, KG-Rank innovatively applies multiple ranking techniques to refine the ordering of these triples, providing more relevant and precise information for LLM inference. To the best of our knowledge, KG-Rank is the first application of KG combined with ranking models in medical QA specifically for generating long answers. Evaluation on four selected medical QA datasets demonstrates that KG-Rank achieves an improvement of over 18% in ROUGE-L score. Additionally, we extend KG-Rank to open domains, including law, business, music, and history, where it realizes a 14% improvement in ROUGE-L score, indicating the effectiveness and great 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 generative capabilities (Gao et al., 2023; Yang et al., 2024b). Despite their considerable potential in various domains, including medicine (Li et al., 2022a; Yang et al., 2023c; Ke et al., 2024; Yang et al., 2024a), their limited training on medical data raises concerns about the consistency of the generated con-

tent with established medical facts (Yang et al., 2023b; Bi et al., 2024).

To address this challenge without additional computational cost, previous research, such as Almanac (Hiesinger et al., 2023) and ChatENT (Long et al., 2023), leverages external medical knowledge to enhance the accuracy and reliability of LLM-generated content. However, merely retrieving external knowledge risks introducing irrelevant or unreliable information (Yang et al., 2024a), which can compromise the effectiveness of LLMs, and raise issues of credibility, data consistency, privacy, security, and legality. While previous studies have emphasized the advantages of utilizing external knowledge, they have overlooked a crucial question: *How to better integrate external knowledge?*

In this work, we propose **KG-Rank**, an augmented framework that integrates a structured medical knowledge graph (KG) with ranking techniques into LLMs to achieve more accurate and reliable long-form medical question-answering (QA). We first retrieve one-hop relations of related medical entities from the medical KG (Unified Medical Language System (UMLS)) (Bodenreider, 2004). To retain relevant information from the KG, we then propose to apply ranking and re-ranking methods to optimize the ordering of triplets.

Specifically, we introduce three ranking techniques to improve the integration of LLM with KG by filtering irrelevant data, highlighting key information, and ensuring diversity. These techniques also streamline the process by reducing the number of triplets required for LLM inference. Additionally, we apply re-ranking models to reassess and emphasize the most relevant triplets, enhancing the factuality of KG-Rank in the long-form medical QA task.

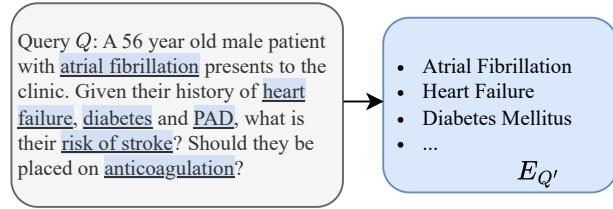
To summarize, our contributions are: (1) We propose KG-Rank, a KG-augmented LLM framework for the medical QA task. To the best of our knowledge, this is the first application of KG com-

bined with ranking techniques to enhance LLMs for medical QA with long answers. (2) We incorporate different ranking and re-ranking techniques to eliminate noise and redundancy in the KG-retrieval stage. (3) We validate the effectiveness of KG-Rank on both medical and various open-domain QA tasks. All the data and code can be found at <https://github.com/YangRui525/KG-Rank>.

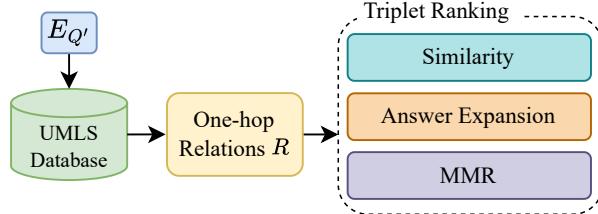
2 Methodology

As shown in Fig. 1, we introduce the KG-Rank (Knowledge Graph -Rank) framework for the long-form medical QA task.

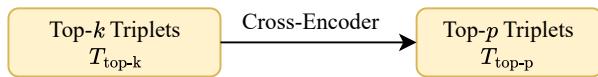
Step 1: Entity Extraction and Mapping



Step 2: Relation Retrieval and Triplet Ranking



Step 3: Re-Ranking



Step 4: Obtaining LLM Response



Figure 1: An illustration of KG-Rank Framework.

2.1 External Knowledge Graph

We define the external KG as $G = (V, E)$, where V represents the set of entities and E represents the set of structural relations. For the medical QA task, we choose UMLS as the primary medical KG. UMLS is a comprehensive repository of health and biomedical vocabularies, designed to promote information standardization and interoperability. The core component of UMLS, the 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. In UMLS, knowledge is represented in the form of triples, which consist of two medical concepts and the relation between them. For example, in the triple (*Myopia, clinically_associated_with, HYPERGLYCEMIA*), "Myopia" and "HYPERGLYCEMIA" are medical concepts, while "clinically_associated_with" is the relation between them.

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_Q , and then we map each entity $e_i \in E_Q$ to the corresponding entity in the knowledge graph G . The detailed prompt can be found in Appendix A.1.

2.3 Relation Retrieval and Triplet Ranking

After identifying the corresponding entities $E_{Q'}$, we retrieve their one-hop relations from the KG (denoted as *UMLS*):

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

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. The detailed prompt for answer expansion can be found in Appendix A.2.

MMR 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, MedCPT (Jin et al., 2023), 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. The detailed prompt can be found in Appendix A.3.

3 Experiments

We conduct experiments on four selected medical QA datasets, in which the answers are free-text, as shown in Tab. 1. LiveQA (Abacha et al., 2017) consists of health questions submitted by consumers to the National Library of Medicine. 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).

Dataset	Sentence (Q)	Word (Q)	Sentence (A)	Word (A)
LiveQA	1.15	14.76	6.96	141.02
ExpertQA-Bio	1.26	21.69	6.18	184.38
ExpertQA-Med	1.37	22.19	5.96	180.55
MedQA	1.02	7.36	3.38	71.48

Table 1: Statistics on the average number of sentences and words across four medical datasets (Q: Question, A: Answer).

3.1 Results

As shown in Tab. 2, we evaluate GPT-4 and LLaMa2-13b across the following settings: zero-shot (ZS), and three proposed ranking techniques:

Similarity Ranking (Sim), Answer Expansion Ranking (AE), and Maximal Marginal Relevance Ranking (MMR). Also with the Re-ranking (RR), which is on top of the Similarity Ranking.

3.2 Datasets

The results show that incorporating the knowledge graph and ranking techniques 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 does not show steady improvement compared to other ranking techniques. 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.

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 on these datasets in Section 4 (Medical LLM).

4 Ablation Study and Analysis

Medical LLM To further investigate the capability of the medical LLM, we compare the zero-shot performance of LLaMa2-7b and baize-healthcare (Xu et al., 2023) without KG-Rank. 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 comparison results can be found in Appendix B.1.

Re-ranking Models We employ GPT-4 with similarity ranking as the final setting and compare two re-ranking models: the MedCPT cross-encoder model, trained on the extensive PubMed articles, and the Cohere (<https://cohere.com>) re-ranking model, designed for broader domain applications. As shown in Tab. 3, MedCPT steadily outperforms the Cohere re-rank model on all datasets, highlighting the importance of specialized re-rank models

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

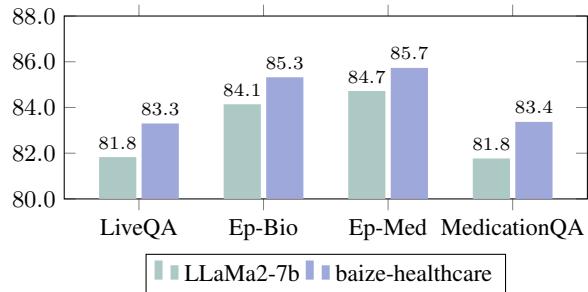


Figure 2: BERTScore comparison: zero-shot setting with LLaMa2-7b and Baize-Healthcare. Ep stands for ExpertQA.

in the medical field. Additional evaluations are provided in Appendix B.2.

Dataset	ROUGE-L	BERTScore	MoverScore	BLEURT
<i>Cohere</i>				
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
<i>MedCPT</i>				
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 3: The performance of Cohere re-rank model and MedCPT in the re-ranking stage.

Case Study To further analyze the generated content of 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. More case studies can be found in the Appendix C.

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.

LLM-based Evaluation Although KG-Rank achieves significant improvements in ROUGE, BERTScore, MoverScore, and BLEURT, these automatic scores may have limitations in evaluating the factuality of long-form medical QA. Therefore, we introduce GPT-4 score specifically for factuality

evaluation (Zheng et al., 2024). The evaluation criteria are designed by two resident physicians with over five years of experience, which can be found in Appendix A.4. As shown in Tab. 4, we choose GPT-4 as the vanilla model, and KG-Rank outperforms the zero-shot setting across all datasets.

Dataset	Zero-Shot	Tie	KG-Rank
LiveQA	0	43	61
ExpertQA-Bio	0	43	52
ExpertQA-Med	3	235	266
MedQA	8	211	468

Table 4: GPT-4 evaluation across four medical datasets.

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 conduct the experiment on Mintaka (Sen et al., 2022), which is a complex, natural, and multilingual dataset designed for experimenting with end-to-end question-answering models. We randomly select 1,000 pairs from the test set for evaluation. Under the enhancement of the KG-Rank framework, the accuracy increases from 60.40% to 61.90%. The detailed prompt can be found in Appendix A.5.

We also conduct experiments in the domains of law, business, music, and history using the ExpertQA dataset. We employ GPT-4 as the vanilla model and use ROUGE-L, BERTScore, and MoverScore for evaluation. As shown in Tab. 5, KG-Rank outperforms the baseline across all benchmarks. Building on these findings, the effectiveness of our framework is not limited to the medical domain but can also be applied to various other fields. For more case studies, please refer to Appendix C.

5 Conclusion

In this work, we propose KG-Rank, an enhanced LLM framework that integrates a medical KG and ranking techniques to improve the factuality of medical QA. As far as we know, KG-Rank is the first application of KG combined with ranking techniques for long-answer medical QA. Across four medical QA datasets, KG-Rank demonstrates over an 18% improvement in ROUGE-L score. Its application to open domains yields a 14% ROUGE-L score enhancement, underscoring KG-Rank’s effectiveness and versatility.

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

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. We will consider graph-based methods (Yang et al., 2023a; Li et al., 2022b) and some efficiency methods (Feng et al., 2023).

Ethical Considerations

This research utilize public medical datasets solely for academic purposes, not for practical application. We employ 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.

- Baolong Bi, Shenghua Liu, Lingrui Mei, Yiwei Wang, Pengliang Ji, and Xueqi Cheng. 2024. **Decoding by contrasting knowledge: Enhancing llms' confidence on edited facts.**
- 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*.
- Aosong Feng, Irene Li, Yuang Jiang, and Rex Ying. 2023. **Diffuser: Efficient transformers with multi-hop attention diffusion for long sequences.** In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 12772–12780. AAAI Press.
- Fan Gao, Hang Jiang, Moritz Blum, Jinghui Lu, Yuang Jiang, and Irene Li. 2023. Large language models on wikipedia-style survey generation: an evaluation in nlp concepts. *arXiv preprint arXiv:2308.10410*.
- 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.
- Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. Medept: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651.
- 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. 2022a. Neural natural language processing for unstructured data in electronic health records: A review. *Computer Science Review*, 46:100511.
- Irene Li, Linfeng Song, Kun Xu, and Dong Yu. 2022b. **Variational graph autoencoding as cheap supervision for AMR coreference resolution.** In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2790–2800. Association for Computational Linguistics.
- 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, Prajwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- 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*.
- Boming Yang, Dairui Liu, Toyotaro Suzumura, Ruihai Dong, and Irene Li. 2023a. **10024 going beyond local: Global graph-enhanced personalized news recommendations.** In *Proceedings of the 17th ACM Conference on Recommender Systems, RecSys 2023, Singapore, Singapore, September 18-22, 2023*, pages 24–34. ACM.
- Rui Yang, Yilin Ning, Emilia Keppo, Mingxuan Liu, Chuan Hong, Danielle S Bitterman, Jasmine Chiat Ling Ong, Daniel Shu Wei Ting, and Nan Liu. 2024a. Retrieval-augmented generation for generative artificial intelligence in medicine. *arXiv preprint arXiv:2406.12449*.
- Rui Yang, Ting Fang Tan, Wei Lu, Arun James Thirunavukarasu, Daniel Shu Wei Ting, and Nan Liu. 2023b. Large language models in health care: Development, applications, and challenges. *Health Care Science*.

Rui Yang, Boming Yang, Sixun Ouyang, Tianwei She, Aosong Feng, Yuang Jiang, Freddy Lecue, Jinghui Lu, and Irene Li. 2024b. Leveraging large language models for concept graph recovery and question answering in nlp education. *arXiv preprint arXiv:2402.14293*.

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. 2023c. [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*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

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 technique.

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 Physician-Designed Criteria for GPT-4 Evaluation

Tab. 6 shows the criteria for evaluating medical long-form QA established by two resident physicians with over five years of experience. This criteria is part of the GPT-4 evaluation prompt.

Evaluation Criteria
Factuality: The degree to which the generated text aligns with established medical facts, providing accurate explanations for further verification.
Readability: The extent to which the generated text is readily comprehensible to the user, incorporating suitable language and structure to facilitate accessibility.
Relevance: The extent to which the generated text directly addresses medical questions while encompassing a comprehensive range of pertinent information.
Completeness: The degree to which the generated text comprehensively portrays the clinical scenario or posed question, including other pertinent considerations.

Table 6: Physician-designed criteria for GPT-4 evaluation.

A.5 KG-Enhanced Prompt for Mintaka Task

Fig. 7 presents the prompt for obtaining KG-enhanced LLM answers, specially designed for the Mintaka 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 performance of widely-used LLMs on four medical datasets under the zero-shot setting. As shown in Tab. 7, the results indicate that GPT-4 performing better than the other LLMs.

Dataset	Evaluation Metrics					
	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	MoverScore	BLEURT
<i>LLaMa2-7b</i>						
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
<i>baize-healthcare</i>						
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
<i>LLaMa2-13b</i>						
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
<i>GPT-4</i>						
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 7: 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 are highlighted in **bold**.

B.2 Performance of Different Re-rank Models

In this section, we evaluate the performance of MedCPT and the Cohere re-rank model on four medical datasets within the GPT-4 with similarity ranking setting. As shown in Table 8, the results indicate that MedCPT outperforms the Cohere re-rank model.

Dataset	GPT-4					
	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	MoverScore	BLEURT
<i>Cohere</i>						
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
<i>MedCPT</i>						
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 8: 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 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.

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.

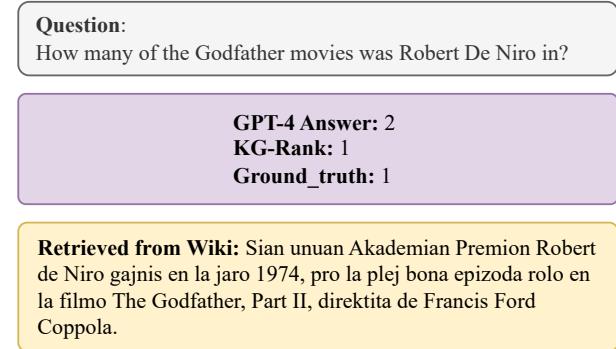


Figure 9: A case study from Mintaka: we show results from vanilla GPT-4 and KG-Rank-enhanced GPT-4.

D Experimental Setup

In our experimental setup, we employ UmlsBERT¹, baize-healthcare², 1llama-2-7b-chat-hf³, 1llama-2-13b-chat-hf⁴, MedCPT⁵ 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. In the MMR Ranking setting, the default value for w is 0.1, and δ is set to 0.01. 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](https://github.com/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>