MKG-Rank: Enhancing Large Language Models with Knowledge Graph for Multilingual Medical Question Answering

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have shown remarkable progress in medical question answering (QA), yet their effectiveness remains predominantly limited to English due to imbalanced multilingual training data and scarce medical resources for low-resource languages. To address this critical language gap in medical QA, we propose **M**ultilingual **K**nowledge **G**raph-based Retrieval **Rank**ing (**MKG-Rank**), a knowledge graph-enhanced framework for multilingual medical QA with English-centric LLMs. It extracts key medical entities from input and translates them at the word level to query the external UMLS knowledge graphs. A multi-angle ranking mechanism filters the retrieved triplets, which are converted into fact statements and inserted into the English-trained LLM, delivering low-cost and accurate medical QA in multiple languages. Extensive experiments on four benchmarks—Chinese, Japanese, Korean, and Swahili—show that MKG-Rank consistently surpasses zero-shot baselines by up to 35.03%, and offers the possibility of supporting privacy-sensitive medical QA via locally deployable alternatives to commercial LLMs. Case studies further demonstrate MKG-Rank surfaces retrieved facts with each answer, providing transparent evidence and paving the way for explainable multilingual medical QA. [1]

1 Introduction

Large Language Models (LLMs) (Hurst et al., 2024; Anthropic, 2024b; Dubey et al., 2024) have achieved remarkable performance in a wide range of Natural Language Processing (NLP) tasks, including question answering (Jiang et al., 2021; Dong et al., 2022) and information retrieval (Wang et al., 2023; Fan et al., 2024). Beyond general NLP applications, LLMs have also been successfully applied to specific professional domains such as medicine and law, demonstrating promising results (Yang et al., 2024c,d; Ke et al., 2024; Zakka et al., 2024; Bernsohn et al., 2024).

While these advances have been demonstrated predominantly in English-language settings, research on their effectiveness in other languages remains relatively underexplored, especially in medical question-answering (Singh et al., 2024a). Specifically, the remaining challenges are: (1) mainstream LLMs are predominantly trained with English-centric data, resulting in a highly unbalanced distribution between languages (Chataigner et al., 2024), which limits their effectiveness in multilingual contexts; and (2) high-quality external medical data for low-resource languages are extremely scarce (Quercia et al., 2024). As a result, existing LLMs exhibit significant performance gaps in multilingual medical applications, limiting their use in non-English-speaking medical settings.

Existing Works and Limitations. Existing methods have emerged but still face significant limitations. *Full-text Translation-based methods* either translate inputs into English for inference (Asai et al., 2018; Montero et al., 2022) or convert rich English corpora into target languages to generate training data (Jundi and Lapesa, 2022; Zhang et al., 2023), both of which incur substantial translation costs and risk semantic distortion or outright inaccuracies in medical content. Alternatively, *data-intensive adaptation techniques* (Yang et al., 2023; Lai et al., 2023; Li et al., 2023a; Üstün et al., 2024) rely on massive multilingual corpora, which are scarce in specialized medical domains. While recent *multilingual RAG systems* (Chirkova et al., 2024; Yang et al., 2024b; Park and Lee, 2025) avoid the need for retraining, they still depend on external multilingual databases, which are scarce or unavailable in low-resource medical contexts.

Our Approach. We propose **M**ultilingual

[1] Code: https://anonymous.4open.science/r/MKG-Rank-6B72. A demo video is in the .zip file with the paper submission.

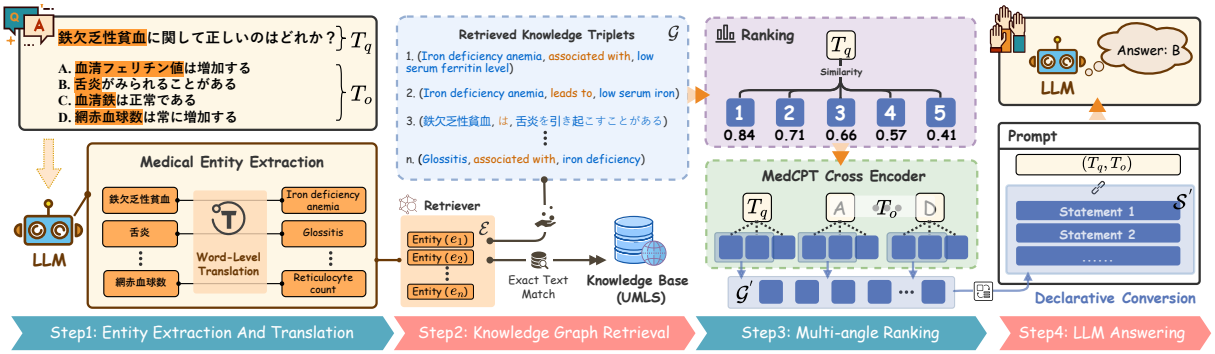

Figure 1: The overall architecture of our proposed MKG-Rank. The English translation of the question and options in the figure is provided in the Appendix H.

Knowledge **G**raph-based Retrieval **Rank**ing (**MKG-Rank**). The system first extracts the salient medical terms from each query and translates only those terms into English, cutting translation cost and avoiding the semantic drift common in full-text translation. It then retrieves relevant entries from an English-centric medical knowledge graph, enabling an English-trained LLM to deliver accurate multilingual medical QA with minimal overhead.

In summary, our contributions are: (1) We propose MKG-Rank, an efficient framework that enables English-centric LLMs to handle multilingual medical QA via word-level translation of key terms to query easily accessible external medical knowledge, reducing semantic drift and computational cost; (2) we conduct extensive experiments on four multilingual medical QA datasets, showing that MKG-Rank consistently outperforms zero-shot base LLMs, with accuracy improvements of up to 35.03%, and also provides reliable local alternatives to commercial LLMs, enhancing privacy in medical scenarios; and (3) our case study shows that MKG-Rank makes its retrieved knowledge explicit, providing clear supporting evidence and paving the way for trustworthy, explainable multilingual medical QA.

2 Methodology

In this section, we introduce our MKG-Rank, as illustrated in Figure 1. It consists of four main steps. First, we extract medical entities from the question and options and translate them into English at the word level. Then, these translated terms are then directly used to query an external medical knowledge graphs (UMLS (Bodenreider, 2004)) via exact text matching to retrieve relevant knowledge graphs (KGs). Next, we propose a multi-angle ranking strategy to to filter and select the most pertinent KG triples. Finally, the selected triplets are converted into declarative statements and, together with the original question and options, fed into the LLM for inference. We followed KG-Rank (Yang et al., 2024c) and applied similar empirical configurations in our method. The detailed prompt can be found in Appendix G

2.1 Medical Entity Extraction and Word-Level Translation

Given a medical question T_q and options T_o, we first use an LLM to extract relevant medical entities. The extracted entities are then translated into English using an LLM, forming the set of English medical entities used for retrieval, denoted as $\mathcal{E} = \{\mathbf{e}_i\}_{i=1,\ldots,n}$. In this case, we only perform word-level translation on the extracted key entity texts rather than translating the entire question or options. This approach effectively reduces translation overhead and avoids semantic drift, a common issue in full-text translation.

2.2 External Medical KGs Retrieval

To retrieve external medical knowledge relevant to the question, we use the translated English medical entities $\mathcal{E}$ to query the Unified Medical Language System (UMLS) (Bodenreider, 2004). Specifically, each medical entity $\mathbf{e}_i$ is used as a query string to perform exact word matching against the UMLS knowledge repository. Each query returns relevant knowledge in the form of triplets (h, r, t), where h and t are medical concepts and r is the semantic relation between them. For example, as illustrated in Figure 1, we can obtain a triplet like ("Iron deficiency anemia", "leads to", "low serum iron") from UMLS. We organize the retrieved triplets and represent them as KGs to stress the knowledge structure, defined as G_i. The medical knowledge retrieved for each entity is aggregated to form the final external knowledge set: $\mathcal{G} = \bigcup_{i=1}^{n} G_i$.

Model	JMMLU		CMMLU		SW MMLU		KO MMLU	
	Base	MKG-Rank	Base	MKG-Rank	Base	MKG-Rank	Base	MKG-Rank
Qwen-2.5 72B (Yang et al., 2024a)	74.00	80.22 (+6.22%)	84.54	81.60 (-2.94%)	44.28	50.90 (+6.62%)	67.72	71.86 (+3.14%)
Llama-3.1 70B* (Grattafiori et al., 2024)	43.33	70.00 (+26.67%)	50.00	72.69 (+22.69%)	36.55	62.34 (+25.79%)	32.97	68.00 (+35.03%)
Claude-3.5 haiku (Anthropic, 2024a)	67.11	76.44 (+9.33%)	50.90	63.21 (+12.31%)	40.28	51.03 (+10.75%)	56.55	68.55 (+12.00%)
GPT-4o-mini (OpenAI, 2024b)	77.33	80.88 (+3.55%)	62.08	70.32 (+8.24%)	66.90	72.14 (+5.24%)	71.59	76.69 (+5.10%)
GPT-4o (OpenAI, 2024a)	83.78	**84.44** (+0.66%)	66.59	**81.83** (+15.24%)	75.86	**83.31** (+7.45%)	78.21	**86.34** (+8.13%)

Table 1: Accuracy comparison between our proposed MKG-Rank and the base models on four multilingual datasets: JMMLU (Japanese), CMMLU (Chinese), SW MMLU (Swahili), and KO MMLU (Korean). * indicates the base model on which MKG-Rank achieves the highest performance gain. The best performance is shown in **bold**.

2.3 Multi-Angle Ranking

We propose a Multi-Angle Ranking mechanism with two-stage filtering to identify and prioritize relevant medical knowledge and mitigate noise, as the retrieved $\mathcal{G}$ is often multilingual and may contain irrelevant content. Given the extracted medical knowledge set $\mathcal{G}$ (a collection of retrieved knowledge triplets), we first compute the similarity between each triplet and the question T_q using UMLS-BERT (Michalopoulos et al., 2021) embeddings. Based on these similarity scores, we rank the triplets and select a set of top candidates. In the second stage, we apply the MedCPT cross encoder[2], trained on 255 million query-article pairs from PubMed search logs, to further refine the selection. For each candidate triplet, we use MedCPT to generate embeddings for the question T_q, each option in T_o, and the triplet itself. Then we compute their relevance to select the top-ranked relevant triplets as the final knowledge set $\mathcal{G}'$.

2.4 Declarative Conversion

For the filtered medical knowledge set $\mathcal{G}'$, we use the LLM to convert each triplet into a declarative statement, forming the set of natural-language statements $\mathcal{S}'$. Finally, the question T_q and options T_o, along with the medical knowledge in declarative form $\mathcal{S}'$, are fed into the LLM for reasoning to generate the final answer, represented as: $y = \text{LLM}(T_q, T_o, \mathcal{S}')$, where y represents the answer generated by the LLM, which uses the refined medical knowledge as retrieved evidence to enhance its reasoning in multilingual medical QA.

3 Experiments

3.1 Datasets

To evaluate the effectiveness of MKG-Rank in multilingual medical QA, we conducted experiments on four multiple-choice datasets covering different languages, focusing on medical-related subsets: JMMLU [3] (Japanese), CMMLU (Li et al., 2023b) (Chinese), KO MMLU (Korean), and SW MMLU (Singh et al., 2024b) (Swahili). Further details can be found in Appendix A.

3.2 Results and Analysis

In Table 1, we compare MKG-Rank (LLM backbone) with the baseline LLMs (zero-shot setting). The results show that our method consistently outperforms all base LLMs. Specifically, with Llama 3.1 70B, we achieved over a 20% improvement across all datasets. For large-scale closed-source LLMs, our method achieved the highest gain on Claude 3.5 Haiku, with an average improvement of 11.1% across the four datasets. Additionally, we achieved an average improvement of 5.5% and 7.8% on GPT-4o-mini and GPT-4o, respectively. Notably, open-source LLMs with MKG-Rank outperform GPT in certain cases, suggesting our method could serve as reliable local alternatives to GPT, especially in medical scenarios where privacy is a concern. Interestingly, Qwen 2.5 72B shows a performance drop on CMMLU, which primarily because its strong Chinese training corpus makes English medical knowledge integration interfere with its reasoning. As an extension, we conducted additional experiments on small-scale LLMs below 32B, as shown in Appendix D.

3.3 Ablation Study and Analysis

Effectiveness of the Declarative Conversion. To evaluate the effectiveness of the declarative conversion in Section 2.4, we compare the performance of Qwen-2.5 70B and GPT-4o-mini with and without this mechanism, as shown in Figure 3. The experimental results show that the declarative conversion mechanism significantly improves the performance of the base models, with particularly notable improvements observed on GPT-4o-mini. Directly retrieved knowledge graphs contain multilingual information, which can negatively affect the LLM's

[2] https://huggingface.co/ncbi/MedCPT-Cross-Encoder

[3] https://huggingface.co/datasets/nlp-waseda/JMMLU

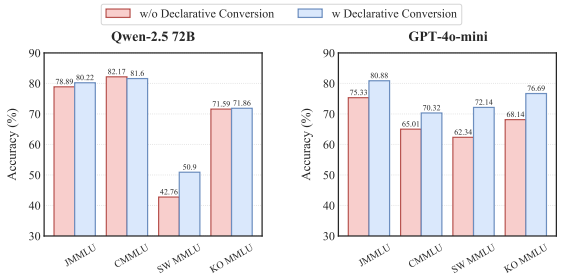

Figure 2: Comparison of the Acc evaluated on Qwen-2.5 72B and GPT-4o-mini across four language datasets with (w/) and without (w/o) declarative conversion.

encoding process, leading to accuracy degradation, especially on the SW MMLU dataset. The proposed declarative conversion mechanism addresses this issue by converting the retrieved raw data, allowing the model to focus on high-relevance English medical concepts.

Performance Evaluation under CoT. We evaluated the effectiveness of MKG-Rank under the CoT (Chain-of-Thought) setting, as shown in Figure 2. The experimental results show that our method achieves strong performance even under CoT, consistently outperforming the baseline across all four datasets, especially on the SW MMLU and KO MMLU datasets. Additional ablation studies are provided in Appendix E.

3.4 Case Study

We conduct case studies on both Japanese and Chinese scenarios, as shown in Figure 4. For scenario 1, we first extract relevant medical entities from the given question & options and translate them into English (e.g., *diplopia*, *fourth nerve palsy*) for querying external medical knowledge graphs UMLS. The retrieved medical KGs are multilingual and may contain redundant or irrelevant information. Our Multi-Angle Ranking strategy effectively filters out unrelated content. The filtered medical triples are then converted into natural-language statements. Finally, the LLM makes the final decision by reasoning over the original question, options, and the obtained medical knowledge statements. Similarly, in the Chinese QA scenario (Scenario 2), we first extract medical entities from the Chinese input and translate them into English for querying. The retrieved knowledge is also represented as English statements and fed to the LLM for reasoning.

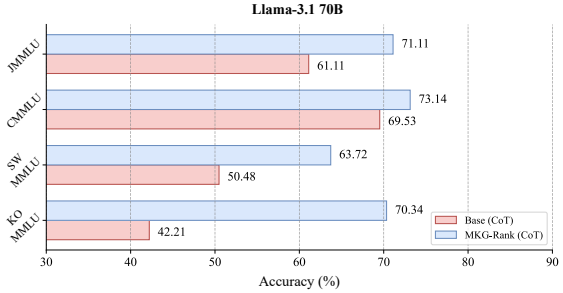

Figure 3: Evaluation on Llama-3.1 70B across four language datasets under CoT.

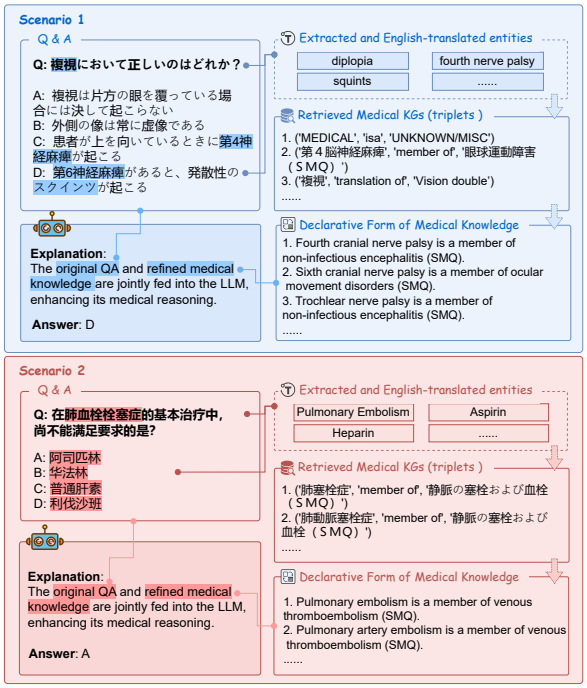

Figure 4: Case Study. More details, along with the English version of the questions and options are provided in the Appendix I.

These case studies demonstrate how MKG-Rank improves transparency and trustworthiness by explicitly presenting supporting evidence during the reasoning process.

4 Conclusion

In this work, we propose MKG-Rank, a knowledge graph-augmented framework that enables English-centric LLMs to effectively handle multilingual medical QA. By leveraging comprehensive external medical knowledge graphs and introducing a word-level translation mechanism, MKG-Rank effectively bridges the medical knowledge gap between English and other languages. Furthermore, we design a Multi-angle ranking mechanism to filter relevant results, ensuring more accurate answers. Extensive experiments across four languages demonstrate that MKG-Rank consistently outperforms zero-shot LLM baselines in multilingual medical QA scenarios.

Limitations

In this study, we developed an enhanced framework based on KG-Rank (Yang et al., 2024c) to improve the performance of LLMs in medical question answering. However, this framework also has certain limitations in practical applications, which we will discuss in the next phase. For incremental databases, it is necessary to set a time for retrieval from the cloud to achieve a balance between efficiency and effectiveness. In the future, we plan to explore a reinforcement learning approach (Chen et al., 2025) to strike a balance between exploitation and exploration, optimizing the reasoning process while leveraging the model's inherent knowledge.

References

Anthropic. 2024a. Claude 3.5 haiku. https://www.anthropic.com/claude/haiku. Retrieved March 13, 2025.

Anthropic. 2024b. Claude 3.5 sonnet. https://www.anthropic.com/news/claude-3-5-sonnet. Retrieved March 13, 2025.

Akari Asai, Akiko Eriguchi, Kazuma Hashimoto, and Yoshimasa Tsuruoka. 2018. Multilingual extractive reading comprehension by runtime machine translation. *arXiv preprint arXiv:1809.03275*.

Dor Bernsohn, Gil Semo, Yaron Vazana, Gila Hayat, Ben Hagag, Joel Niklaus, Rohit Saha, and Kyryl Truskovskyi. 2024. Legallens: Leveraging llms for legal violation identification in unstructured text. In *Conference of the European Chapter of the Association for Computational Linguistics*.

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

Cléa Chataigner, Afaf Taïk, and Golnoosh Farnadi. 2024. Multilingual hallucination gaps in large language models. *arXiv preprint arXiv:2410.18270*.

Yiqun Chen, Lingyong Yan, Weiwei Sun, Xinyu Ma, Yi Zhang, Shuaiqiang Wang, Dawei Yin, Yiming Yang, and Jiaxin Mao. 2025. Improving retrieval-augmented generation through multi-agent reinforcement learning. *arXiv preprint arXiv:2501.15228*.

Nadezhda Chirkova, David Rau, Hervé Déjean, Thibault Formal, Stéphane Clinchant, and Vassilina Nikoulina. 2024. Retrieval-augmented generation in multilingual settings. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 177–188, Bangkok, Thailand. Association for Computational Linguistics.

Xiangjue Dong, Jiaying Lu, Jianling Wang, and James Caverlee. 2022. Closed-book question generation via contrastive learning. *arXiv preprint arXiv:2210.06781*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Wenqi Fan, Yujuan Ding, Liang bo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Knowledge Discovery and Data Mining*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Iman Jundi and Gabriella Lapesa. 2022. How to translate your samples and choose your shots? analyzing translate-train & few-shot cross-lingual transfer. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 129–150, Seattle, United States. Association for Computational Linguistics.

Yuhe Ke, Rui Yang, Sui An Lie, Taylor Xin Yi Lim, Yilin Ning, Irene Li, Hairil Rizal Abdullah, Daniel Shu Wei Ting, and Nan Liu. 2024. Mitigating cognitive biases in clinical decision-making through multi-agent conversations using large language models: simulation study. *Journal of Medical Internet Research*, 26:e59439.

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. *arXiv preprint arXiv:2307.16039*.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023a. Bactrian-x:

Multilingual replicable instruction-following models with low-rank adaptation. *arXiv preprint arXiv:2305.15011*.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023b. Cmmlu: Measuring massive multitask language understanding in chinese. *arXiv preprint arXiv:2306.09212*.

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. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1744–1753, Online. Association for Computational Linguistics.

Ivan Montero, Shayne Longpre, Ni Lao, Andrew Frank, and Christopher DuBois. 2022. Pivot through English: Reliably answering multilingual questions without document retrieval. In *Proceedings of the Workshop on Multilingual Information Access (MIA)*, pages 16–28, Seattle, USA. Association for Computational Linguistics.

OpenAI. 2024a. Gpt-4o. `https://openai.com/index/hello-gpt-4o/`. Retrieved March 13, 2025.

OpenAI. 2024b. Gpt-4o-mini. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/. Retrieved March 13, 2025.

Jeonghyun Park and Hwanhee Lee. 2025. Investigating language preference of multilingual rag systems. *arXiv preprint arXiv:2502.11175*.

A. Quercia, Jamil Zaghir, Christian Lovis, and Christophe Gaudet-Blavignac. 2024. Medfrenchmark, a small set for benchmarking generative llms in medical french. *Studies in health technology and informatics*, 316:601–605.

Abhishek Kumar Singh, Vishwajeet Kumar, Rudra Murthy, Jaydeep Sen, Ashish Mittal, and Ganesh Ramakrishnan. 2024a. Indic qa benchmark: A multilingual benchmark to evaluate question answering capability of llms for indic languages. *ArXiv*, abs/2407.13522.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, et al. 2024b. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *arXiv preprint arXiv:2412.03304*.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, et al. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

Keheng Wang, Feiyu Duan, Sirui Wang, Peiguang Li, Yunsen Xian, Chuantao Yin, Wenge Rong, and Zhang Xiong. 2023. Knowledge-driven cot: Exploring faithful reasoning in llms for knowledge-intensive question answering. *ArXiv*, abs/2308.13259.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024a. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Eugene Yang, Thomas Jänich, James Mayfield, and Dawn Lawrie. 2024b. Language fairness in multilingual information retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2487–2491.

Rui Yang, Haoran Liu, Edison Marrese-Taylor, Qingcheng Zeng, Yu He Ke, Wanxin Li, Lechao Cheng, Qingyu Chen, James Caverlee, Yutaka Matsuo, et al. 2024c. Kg-rank: Enhancing large language models for medical qa with knowledge graphs and ranking techniques. *arXiv preprint arXiv:2403.05881*.

Rui Yang, Qingcheng Zeng, Keen You, Yujie Qiao, Lucas Huang, Chia-Chun Hsieh, Benjamin Rosand, Jeremy Goldwasser, Amisha Dave, Tiarnan Keenan, et al. 2024d. Ascle—a python natural language processing toolkit for medical text generation: development and evaluation study. *Journal of Medical Internet Research*, 26:e60601.

Wen Yang, Chong Li, Jiajun Zhang, and Chengqing Zong. 2023. Bigtranslate: Augmenting large language models with multilingual translation capability over 100 languages. *arXiv preprint arXiv:2305.18098*.

Cyril Zakka, Rohan Shad, Akash Chaurasia, Alex R Dalal, Jennifer L Kim, Michael Moor, Robyn Fong, Curran Phillips, Kevin Alexander, Euan Ashley, et al. 2024. Almanac—retrieval-augmented language models for clinical medicine. *Nejm ai*, 1(2):AIoa2300068.

Ge Zhang, Yemin Shi, Ruibo Liu, Ruibin Yuan, Yizhi Li, Siwei Dong, Yu Shu, Zhaoqun Li, Zekun Wang, Chenghua Lin, et al. 2023. Chinese open instruction generalist: A preliminary release. *arXiv preprint arXiv:2304.07987*.

A Dataset Details

JMMLU [4]. It consists of a subset of Japanese-translated questions from MMLU (Hendrycks et al., 2020) and questions based on the Japanese cultural context. We selected three medically related subsets, which contain 450 entries.

CMMLU (Li et al., 2023b). It is a multi-task benchmark designed for Chinese language understanding, consisting of multiple-choice questions with four options.

KO MMLU, SW MMLU. Global MMLU (Singh et al., 2024b) is a multilingual version derived from MMLU, which includes carefully translated and machine-translated versions in various languages. We select Korean and Swahili (a language widely spoken in East Africa) from this dataset, referred to as KO MMLU and SW MMLU, respectively.

From the aforementioned datasets, we select the medically related subsets, which include Clinical Knowledge, Professional Medicine, and College Medicine. More information is shown in Table 2.

Dataset	Language	Size	Length
JMMLU	Japanese	450	171
CMMLU	Chinese	886	70
SW MMLU	Korean	725	511
KO MMLU	Swahili	725	215

Table 2: Statistics of evaluation datasets, including the size of each dataset and the average text length of each question and its corresponding options.

B Evaluation Metric

We use accuracy (Acc) as the evaluation metric, which measures the percentage of correct answers provided by the model. Furthermore, any response expressing uncertainty or listing multiple candidate answers is considered incorrect.

C Resource Consumption

	JMMLU 450		CMMLU		SW MMLU		KO MMLU	
Model	A100(hours)	API($)	A100(hours)	API($)	A100(hours)	API($)	A100(hours)	API($)
Qwen-2.5 72B	12	0.08	18	0.22	16	0.17	16	0.17
Llama-3.1 70B	12	0.08	18	0.22	16	0.17	16	0.17
Claude-3.5 haiku	-	1.84	-	3.6	-	2.97	-	2.97
GPT-4o-mini	-	0.26	-	0.54	-	0.44	-	0.44
GPT-4o	-	1.75	-	3.44	-	2.82	-	2.82

Table 3: Resources consumed in the relevant experiments.

D Additional Evaluation Experiments on Small-scale LLMs

To further demonstrate the effectiveness of MKG-Rank, we evaluate its performance against small-scale baseline LLMs on the JMMLU dataset, as shown in Table 4. Experimental results show that MKG-Rank consistently outperforms all small-scale baseline LLMs.

E Additional Ablation Study on Declarative Conversion

We conducted an additional ablation study on Declarative Conversion across three LLMs, as shown in Figure 5. Notably, on the Llama-3.1 70B model, Declarative Conversion demonstrates a negative impact,

[4] https://huggingface.co/datasets/nlp-waseda/JMMLU

Method	Borea-Phi-3.5	Llama-3.2 3B	Qwen-2.5 7B	Meta-Llama-3.1 8B	Qwen-2.5 14B	Phi4 14B	Qwen-2.5 32B
Base	42.00	36.10	58.40	48.22	70.88	67.56	75.11
MKG-Rank	43.43 (+1.43%)	39.78 (+3.68%)	61.33 (+2.93%)	52.22 (+4.00%)	73.11 (+2.23%)	79.56 (+12.00%)	76.00 (+0.89%)

Table 4: Accuracy comparison between MKG-Rank and small-scale base models on the JMMLU dataset.

leading to a decline in performance. After analyzing the results, we believe that the longer transcription context, compared to directly using triples, imposes a greater reasoning burden than the multilingual effect. Specifically, the length of the context appears to affect the LLaMA model's performance more significantly.

F Additional Experiment on Few-shot Prompting

We conducted an additional experiments on the Llama-3.1 70B using few-shot prompting on JMMLU dataset, as shown in Figure 6. The results demonstrate that the improvement brought by MKG-Rank under the few-shot prompting setting is limited. This may be because the in-context samples already provide sufficient knowledge, reducing the effectiveness of retrieving external medical knowledge.

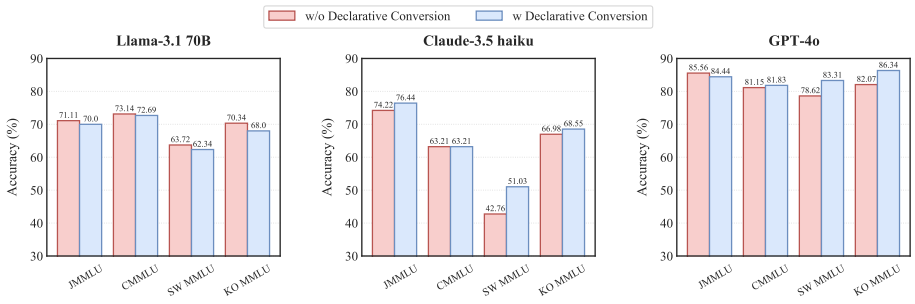

Figure 5: Additional ablation experiments on Llama 70B, Claude-3.5 haiku, and GPT-4o across four language datasets with (w) and without (w/o) multi-angle ranking.

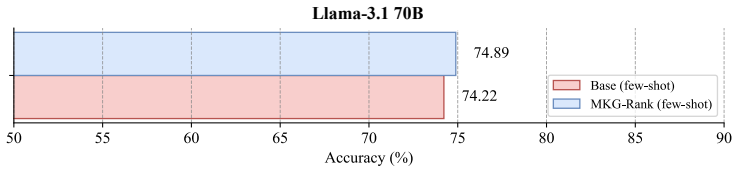

Figure 6: Additional experiment on Llama-3.1 70B using fewshot prompting on JMMLU dataset.

G Prompts

In this section, we will present the prompts used in each stage of reasoning within the MKG-Rank.

G.1 Medical NER Prompt

Figure 7 and Figure 8 illustrate the prompt designed for extracting medical entities from both questions and options, with different extraction counts set for questions and options respectively.

G.2 Declarative Conversion

Figure 9 illustrates the prompt designed for declarative conversion.

G.3 MKG-Rank Enhanced Reasoning Prompt

Figure 10 illustrates the prompt designed for reasoning based on the final integrated knowledge.

G.4 MKG-Rank Enhanced Reasoning Prompt with CoT

Figure 11 illustrates the CoT prompt designed for reasoning based on the final integrated knowledge.

G.5 MKG-Rank Enhanced Reasoning Prompt with Few-shot

Figure 12 illustrates the Few-shot prompt designed for reasoning based on the final integrated knowledge.

---

text:{**question**}

Please extract no more than three medical terminologies that you think are important and related to medical entities from the provided text, and it is not required to be general entity words. Only the corresponding results are returned in json format, and no additional explanation is needed.

-- Examples of results:
{"medical entities" : ["term1", "term2", ...]}

result:

Figure 7: Prompt for extracting medical entities from question.

---

text:{**options**}

Please extract 1 medical term each from the options provided. It should not be a general entity word. Only the corresponding results are returned, and no other explanation is needed.

-- Examples of results:
{"medical entities" : ["term1", "term2", ...]}

result:

Figure 8: Prompt for extracting medical entities from options.

You are an intelligent assistant in the medical field.
Convert all background knowledge into English declarative sentences. Anything you don't think is medically relevant can be deleted.

- Background Knowledge: {**triples**}

Converted Background Knowledge:

Figure 9: Prompt for declarative conversion.

H English Translation of the Question and Options in Figure 1

We provide the English translation of the question and options in Figure 1 for clearer description, as shown in Figure 13.

I A Detailed Case Study with English Annotations

We provide detailed information on the cases in Figure 4, along with the English version of the questions and options, as shown in Figure 14.

One of the following four options is correct. Please choose the option corresponding to the correct answer according to the background knowledge provided and your own knowledge.

You can try to answer the questions in English.

If you think the question is logical, think it step by step, but you only need to return the option letter corresponding to the final result.

- Question: {**question**}

- Options: {**options**}

- Background Knowledge: {**background_knowledge**}

- Answer:
[your option]

Figure 10: Prompt for MKG-Rank enhanced reasoning.

One of the following four options is correct. Please choose the option corresponding to the correct answer according to the background knowledge provided and your own knowledge.

- Question: {**question**}

- Options: {**options**}

- Background Knowledge: {**background_knowledge**}

Let's think step by step.

Format:
Thinking steps:
Step 1: Analyze each option in detail.
Step 2: Consider the likelihood of each option being correct.
Step 3: Apply relevant medical knowledge.
Step 4: Choose the best answer based on reasoning.

Answer: [A/B/C/D]

- Answer:
[your option]

Figure 11: CoT Prompt for MKG-Rank enhanced reasoning.

One of the following four options is correct. Please choose the option corresponding to the correct answer according to the background knowledge provided and your own knowledge.

- Question: {**question**}

- Options: {**options**}

- Background Knowledge: {**background_knowledge**}

- Answer:
[your option]

Here are some examples of correct answers:

Example 1:
Question: Which of the following is the most common cause of chronic kidney disease?
Options:
A) Diabetes mellitus
B) Hypertension
C) Glomerulonephritis
D) Polycystic kidney disease
Correct Answer: A

Example 2:
Question: The most common symptom of myocardial infarction is:
Options:
A) Shortness of breath
B) Chest pain
C) Nausea
D) Diaphoresis
Correct Answer: B

Figure 12: Few-shot Prompt for MKG-Rank enhanced reasoning.

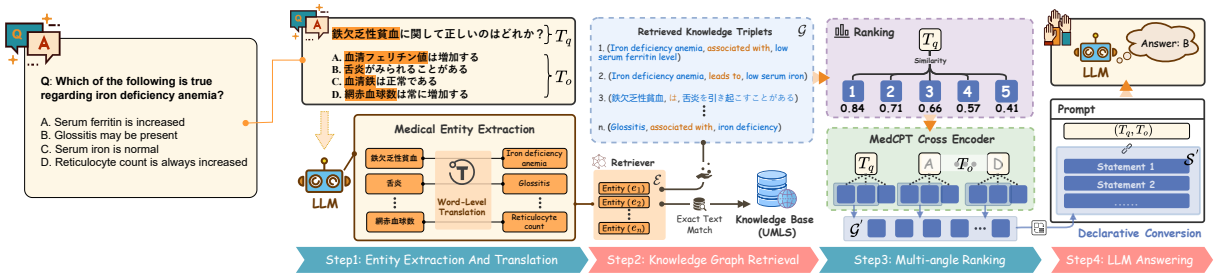

Figure 13: The English version of the question and options in Figure 1.

Figure 14: A detailed case study with comprehensive information, including the English version of the questions and options.