A Pilot Empirical Study on When and How to Use Knowledge Graphs as Retrieval Augmentation Generation

Anonymous ACL submission

Abstract

The integration of Knowledge Graphs (KGs) into the Retrieval Augmentation Generation (RAG) framework has attracted significant interest, with early studies showing promise in mitigating hallucinations and improving model accuracy. However, a systematic understanding and comparative analysis of the rapidly emerging KG-RAG methods are still lacking. This paper seeks to lay the foundation for systematically answering the question of *when and how to use KG-RAG* by analyzing their performance in various application scenarios associated with different technical configurations. After outlining the mind map using KG-RAG framework and summarizing its popular pipeline, we conduct a pilot empirical study of KG-RAG works to reimplement and evaluate 6 KG-RAG methods across 7 datasets in diverse scenarios, analyzing the impact of 9 KG-RAG configurations in combination with 17 LLMs. Our results underscore the critical role of appropriate application conditions and optimal configurations of KG-RAG components. The data and methods used, along with our reimplementation, are publicly available on `https://anonymous.4open.science/r/Understanding-KG-RAG-EB54`.

1 Introduction

Recently, Large Language Models (LLMs) have demonstrated remarkable capabilities in Natural Language Processing (NLP) tasks (Wei et al., 2022a; Brown et al., 2020). However, LLMs face critical challenges including hallucination (Sahoo et al., 2024a), limited incorporation with real-time knowledge (Mallen et al., 2023), and opaque reasoning processes (Zhou et al., 2024). Thus, Retrieval-Augmented Generation (RAG) (Guu et al., 2020) frameworks have emerged as a promising solution by searching most relevant contents from external knowledge base using similarity methods (Fan et al., 2024). However, RAG typically treats document contents as independent units, struggling to capture complex relational information and hierarchical interconnections within the data (Liu et al., 2024; Li et al., 2025c).

To address above limitations, graph-based RAG (Edge et al., 2024), particularly those incorporating Knowledge Graphs (KGs) known as KG-RAG, has emerged as a promising paradigm (Zhang et al., 2022; Guan et al., 2024; Kim et al., 2023; Saleh et al., 2024). KG-RAG leverages semantic relationships between entities (Li et al., 2025a) to enable more sophisticated reasoning capabilities (Sun et al., 2023; Wang et al., 2025) and enhance performance in domain-specific applications (Wen et al., 2024).

However, due to the rapid proliferation of related techniques, these KG-RAG works have emerged in a disjointed manner, much like mushrooms after rain, with significant variations in their use of scenarios, datasets, KG-RAG configurations, and LLMs. They tend to focus on isolated technical innovations across different pipeline stages, without systematic comparison across varied scenarios. Moreover, recent reviews (Pan et al., 2024; Zhang et al., 2025; Peng et al., 2024; Zhao et al., 2024) primarily focuses on qualitative analyses, with a lack of quantitative assessments regarding the impact of key configurations across different task scenarios.

To address this research gap, we aim to explore the key factors that answer the questions of *when* and *how* to use KG-RAG, thereby laying the foundation for a quantitative empirical study. Specifically, we identify two critical gaps in current KG-RAG research: its applicability across diverse scenarios and the effectiveness of different pipeline configurations. First, the applicability of KG-RAG remains insufficiently explored across several dimensions: task domains (ranging from open-domain to domain-specific tasks), task difficulty levels (from single-hop to multi-hop questions) (Zhao et al., 2024), LLM capabilities (from open-source to commercial models), and KG quality (from specialized to general KGs). Second, the impact of different KG-RAG configurations lacks systematic understanding: (1) pre-retrieval query enhancement strategies (query expansion, decomposition, and understanding), (2) varying retrieval forms (from facts to paths and subgraphs), and (3) post-retrieval prompting approaches (e.g., Chain-of-Thought (Wei et al., 2022b) and Tree-of-Thought (Yao et al., 2023)). Through such a systematic investigation, we aim to provide practical guidelines of KG-RAG for answering when and how to use KG-RAG effectively.

In this paper, as a pilot empirical study of the KG-RAG methodology, we reimplement and evaluate 6 KG-RAG methods across 7 datasets in diverse scenarios, analyzing the impact of 9 KG-RAG configurations in combination with 17 LLMs. Our results underscore the crucial role of selecting appropriate application conditions and optimizing the configurations of KG-RAG components. Specifically, we systematically address how much the KG-RAG approach benefits open-source LLMs across different task domains and difficulty levels, and whether these enhancements offer a greater advantage compared to larger or commercial LLMs. Additionally, we examine the influence of various configurations on KG-RAG performance and identify several limitations in current KG-RAG research.

2 Literature Review

Recent surveys and systematic reviews have provided comprehensive analyses of RAG frameworks and their integration with KGs (Pan et al., 2024; Zhao et al., 2024), establishing a solid foundation for understanding this rapidly evolving field. CRAG (Yang et al., 2024c) advances the field by introducing a comprehensive benchmark that evaluates RAG performance across multiple dimensions, including domain specificity, data dynamism, content popularity, and question complexity. Complementary research on RAG optimization strategies (Li et al., 2025b) has investigated the impact of various factors on generation quality, such as model size, prompt design and knowledge base scale. While these studies primarily focus on unstructured text retrieval, their insights provide valuable reference points for understanding structured knowledge retrieval systems like KG-RAG.

The integration of KGs with RAG has attracted significant attention from the research community (Pan et al., 2024). Several comprehensive surveys have systematically documented the evolution, technical frameworks, and key components of KG-RAG (Zhang et al., 2025). These reviews provide extensive coverage of retrieval methods, model architectures, knowledge graph variants, and practical applications (Peng et al., 2024), along with discussions of available open-source implementations and benchmark datasets (Zhao et al., 2024). These surveys primarily focus on taxonomic classification and theoretical analysis, offering valuable qualitative insights into the KG-RAG landscape.

Although existing works demonstrate breadth in their coverage, these studies show deficiencies in quantifying the advantages and disadvantages of different KG-RAG approaches, analyzing their inherent trade-offs, and providing comprehensive experimental data, thus limiting systematic understanding of KG-RAG's effectiveness and optimal configurations across different task scenarios.

3 KG-RAG Scenario and Configuration

As outlined in Sec. 1 and 2, KG-RAG works have emerged in a disjointed manner, with significant variations in the use of scenarios, datasets, KG-RAG configurations, and LLMs. However, current reviews on KG-RAG primarily focuses on qualitative analyses, with a lack of quantitative assessments regarding the impact of key configurations across various task scenarios. To bridge this gap, we explore the key factors that answer the questions of *when* and *how* to use KG-RAG, laying the foundation for a quantitative empirical study.

3.1 When to Use KG-RAG for LLMs?

As discussed in Fig. 1, answering the question of when to use KG-RAG requires considering several factors: 1) whether the task scenario necessitates KG-RAG assistance for the LLM, 2) whether the capabilities of the given LLMs require external knowledge to complete the task, and 3) whether the quality of the KG is sufficient to support the reasoning needs of the LLM.

Task Scenarios. To investigate the applicability of KG-RAG, we categorize task scenarios from two perspectives: task domain and task difficulty.

- Task Domain: Inspired by CRAG (Yang et al., 2024c), we roughly categorize tasks in existing KG-RAG works into open-domain question answering (QA), domain-specific QA and exam. The open-domain QA require general world knowledge, while domain-specific QA focus on

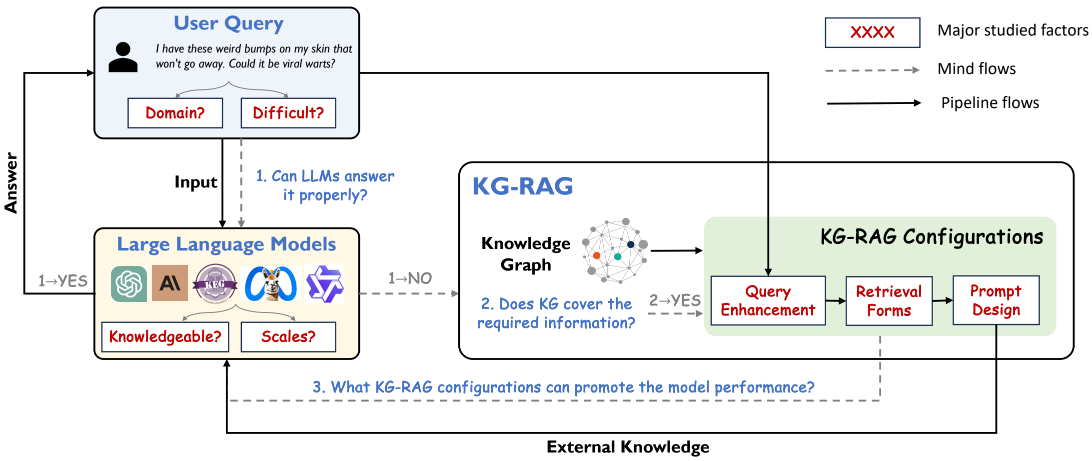

Figure 1: The mind and pipeline flows of KG-RAG.

specialized fields requiring professional knowledge. Domain-specific exam is professional qualification examinations that test domain expertise.

- Task Difficulty: There is currently no clear consensus on how to define task difficulty. After reviewing KG-RAG datasests, we adopt a two-level classification (Zhao et al., 2024). The L1 difficulty involves questions that require straightforward answers based on clear facts (single-hop). The L2 or higher difficulty represent questions that require reasoning and the integration of multiple pieces of information (multi-hop).

Based on the task domain and difficulty, We summarize five representative datasets of KG-RAG works in Table 1. CommonsenseQA (Talmor et al., 2019) is an open-domain QA dataset focusing on commonsense questions. GenMedGPT-5K (Li et al., 2023b) and CMCQA (Xia et al., 2022) are medical consultation datasets for Domain-specific QA, with CMCQA showing higher difficulty through multi-round conversations (more L2 questions). CMB-Exam (Wang et al., 2024a) and ExplainCPE (Li et al., 2023a) are medical professional examination datasets containing both L1 and L2 questions. More detailed information on these datasets can be found in Appx. A.

Capability of LLMs. Beyond the task difficulty, the capability of LLMs is also a key factor in determining the importance of KG-RAG. Considering practical issues such as economics, open-source availability, and data privacy, there is a general hope that open-source LLMs (especially those with low resource consumption) can outperform commercial ones in specialized tasks after incorporating external knowledge. Thus, we include 17 commonly used LLMs of varying scales and types:

- Qwen1.5-7B (Team, 2024) and Llama2-7B (Touvron et al., 2023) serve as the backbone open-source LLMs (BOS-LLMs) for KG-RAG, as they are fully open-source and share comparable architectures (7B, decoder-only).

- Other open-source LLMs: Qwen2.5-7B (Yang et al., 2024a), Qwen2-72B (Yang et al., 2024b), Deepseek-v2-lite (Shao et al., 2024), ChatGLM4-9B (GLM et al., 2024), and Yi-34B (Young et al., 2024) for Chinese; Llama3.2-1B, Llama3-8B (Dubey et al., 2024), Llama2-70B, Gemma2-9B (Team et al., 2024), Mixtral-8*7B (Jiang et al., 2024a) for English; and a domain-specialized model OpenBioLLM-70B (Ankit Pal, 2024).

- Commercial LLMs: Claude3.5-Sonnet, Gemini1.5 - Pro, GPT4o, o1-mini.

Knowledge Graphs. Once the task scenario and LLMs' capabilities are clearly outlined, KG quality will become another decisive factor. Following Wen et al. (2024), we utilize EMCKG and CMCKG as the KGs for GenMedGPT-5 and CMCQA, respectively. Besides, we construct the corresponding KGs for the remaining datasets (detailed in Appx. B). Furthermore, to examine the impact of KG quality (Sui and Hooi, 2024) on KG-RAG, we conducted experiments on the ExplainCPE dataset using spKG (specialized KG) and CMCKG (only partially covers the required knowledge) in Tab. 6.

3.2 How to Use KG-RAG Techniques?

As shown in Fig. 2, to answer the question of how to use KG-RAG, we review five existing KG-RAG works (KGRAG (Soman et al., 2023), ToG (Sun et al., 2023), MindMap (Wen et al., 2024), RoK (Wang et al., 2024b), KGGPT (Kim

Table 1: The statistics of datasets adopted in this paper.

Task Scenario	Dataset	Concrete Task	# Question	Language	# $L1$	# $L2$
Open-domain QA	CommonsenseQA	Commonsense QA	700	English	100%	-
Domain-specific QA	GenMedGPT-5K	Diagnosis	700	English	25.7%	74.3%
	CMCQA	Diagnosis	500	Chinese	-	100%
Domain-specific Exam	CMB-Exam	Multi-choice	3,000	Chinese	74.4%	25.6%
	ExplainCPE	Multi-choice	507	Chinese	49.3%	50.7%

et al., 2023)) and summarize three main modules based on the retrieval stage: Pre-Retrieval, Retrieval, and Post-Retrieval. Additionally, to facilitate subsequent ablation experiments for validating modules, we supplement a experimental Pilot method, as proposed in this paper.

Query Enhancement in Pre-retrieval. The Pre-Retrieval phase focuses on determining "what to retrieve" by aligning queries with knowledge base content (Jiang et al., 2024b). We examine three distinct approaches to query enhancement:

- Query Expansion: RoK (Wang et al., 2024b) leverages Chain-of-Thought (Wei et al., 2022b) to extract key entities through step-by-step reasoning first, enabling the discovery of more relevant entities during retrieval by aligning LLMs' pre-trained knowledge with knowledge in KGs.

- Query Decomposition: KGGPT (Kim et al., 2023) addresses multi-hop reasoning by breaking down complex queries into simpler clauses, making it easier to construct evidence graphs through separate retrievals for each clause.

- Query Understanding: We further integrate query understanding into Pilot, which extracts main ideas from queries using LLMs. It ensures retrieved content aligns with both query and topic, addressing cases where query similarity alone may lead to irrelevant matches (Gan et al., 2024).

Retrieval Forms After Retrieval. In the retrieval phase, KG-RAG organizes retrieved graph context that can be input to LLMs as reference information. Due to differences in specific retrieval mechanisms, the graph context may ultimately be organized into three forms with increasing information granularity: fact, path, and subgraph.

- Fact is the most basic knowledge unit in triplet form `(Subject,Predicate,Object)`, providing discrete, structured knowledge points (Soman et al., 2023). The facts, while precise and processable, lack contextual connections.

- Path consist of connected triplet sequences, offering richer context through interconnected knowledge. ToG (Sun et al., 2023) demonstrates how path-based retrieval supports multi-hop reasoning by guiding LLMs to explore multiple reasoning paths. Paths can balance information density with structural clarity but may miss broader relationships outside the path.

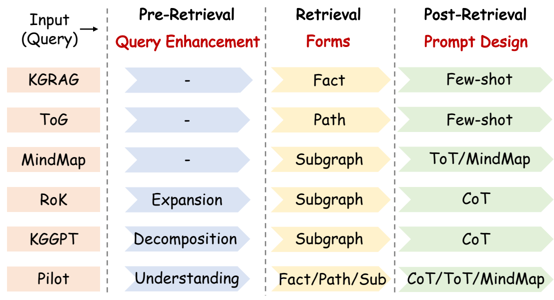

Figure 2: Configurations of KG-RAG

- Subgraph combines both paths and neighboring entity information, can capture more comprehensive relationships and patterns, enabling KG-RAG to understand content more thoroughly and in greater detail. MindMap (Wen et al., 2024) employs both path-based and neighbor-based exploration, ultimately combining path and neighbor information to form an evidence subgraph.

Post-Retrieval: Prompt design. In the Post-Retrieval phase, while some works focus on filtering (Li et al., 2024) or reranking (Glass et al., 2022) retrieved results, we primarily investigate how different prompt designs guide LLMs' reasoning process with retrieved knowledge (Sahoo et al., 2024b; Tonmoy et al., 2024; Chen et al., 2023). We mainly examine three following prompt patterns:

- Chain-of-Thought (CoT) introduces step-by-step reasoning (Wei et al., 2022b), breaking complex problems into sequential intermediate steps.

- Tree-of-Thought (ToT) (Yao et al., 2023) extends this concept by enabling multi-branch exploration, allowing LLMs to simultaneously consider and compare multiple reasoning paths.

- MindMap (Wen et al., 2024) enhances reasoning interpretability by guiding LLMs to construct structured mind maps that integrate retrieved knowledge while maintaining reasoning traces.

Table 2: CommonsenseQA (Self-Construted KG)

Type	Method	Correct	Wrong	Fail
LLM only	Llama3.2-1B	52.93	47.07	0.00
	Llama2-7B	39.06	60.37	0.57
	Llama3-8B	73.82	26.04	0.14
	Llama2-70B	68.1	30.62	1.29
	Mixtral-8*7B	68.53	30.76	0.72
	Gemma-9B	78.83	21.03	0.14
	GPT4o	**84.55**	15.45	0.00
	o1-mini	81.40	18.45	0.14
	Claude3.5-S	82.55	17.45	0.00
	Gemini1.5-P	83.83	16.17	0.00
KG-RAG (Llama2-7B)	KGRAG	42.49	56.94	0.57
	ToG	42.06	57.37	0.57
	MindMap	51.07	47.50	1.43
	RoK	42.86	57.14	0.00
	KGGPT	48.11	51.73	0.15
	Pilot	51.50	48.50	0.00

Table 3: GenMedGPT-5K (EMCKG)

Type	Method	Prec.	Rec.	F1	R-1	R-L
LLM only	Llama3.2-1B	57.32	63.81	60.25	19.37	11.36
	Llama2-7B	58.79	67.89	62.96	21.02	12.21
	Llama3-8B	57.21	63.09	59.87	20.17	11.60
	Llama2-70B	59.35	68.32	63.46	21.32	12.69
	OBLLM-70B	60.54	68.04	64.02	24.28	13.72
	Mixtral-8*7B	59.33	65.53	62.21	24.38	12.79
	GPT4o	56.76	66.08	61.01	23.32	12.62
	o1-mini	58.42	57.47	57.50	17.32	10.59
	Claude3.5-S	57.01	68.35	61.29	22.37	12.01
	Gemini1.5-P	54.49	66.50	59.87	19.07	10.24
KG-RAG (Llama2 -7B)	KGRAG	56.29	67.09	61.17	16.59	10.03
	ToG	56.50	67.80	61.59	16.93	10.06
	MindMap	64.61	62.72	63.58	27.20	17.33
	RoK	59.41	**71.10**	64.68	23.57	14.09
	KGGPT	56.87	68.07	61.92	18.50	10.93
	Pilot	**65.84**	64.49	**65.09**	**28.49**	**17.85**

Table 4: CMCQA (CMCKG)

Type	Method	Prec.	Rec.	F1	R-1	R-L
LLM only	Qwen1.5-7B	67.61	70.57	69.00	16.75	8.91
	Qwen2.5-7B	67.66	70.32	68.91	14.49	7.88
	Qwen2-72B	67.50	70.35	68.84	14.94	8.17
	Deepseek-v2l	67.72	70.19	68.88	15.34	8.57
	ChatGLM-9B	67.53	70.36	68.86	13.95	7.63
	Yi-34B	67.66	70.40	68.94	15.21	8.34
	OBLLM-70B	67.07	69.35	68.14	3.56	3.46
	GPT4o	66.91	70.79	68.74	15.11	7.88
	o1-mini	66.24	69.07	67.55	11.03	6.08
	Claude3.5-S	**68.24**	**72.38**	**70.18**	**18.90**	10.48
	Gemini1.5-P	67.08	70.86	68.86	12.69	6.53
KG-RAG (Qwen1.5 -7B)	KGRAG	65.65	70.01	67.71	16.45	**10.58**
	ToG	65.52	69.64	67.47	13.89	7.30
	MindMap	64.93	66.14	65.46	13.51	7.83
	RoK	66.19	69.73	67.85	15.29	8.00
	KGGPT	66.77	70.40	68.48	15.13	7.87
	Pilot	66.12	70.48	68.17	13.90	7.33

4 Empirical Study

4.1 Research Questions

As discussed in Sec. 2, past reviews primarily provide a macroscopic and qualitative comparison of the differences and similarities among existing KG-RAG works. Therefore, this paper seeks to answer the following research questions (RQs) by conducting a quantitative analysis of various KG-RAG methods and LLMs across different task scenarios:

- **RQ1** (Sec. 4.3.1): How much do the KG-RAG methods benefit the backbone open-source LLMs (BOS-LLMs) across different task scenarios?
- **RQ2** (Sec. 4.3.2): Do BOS-LLMs enhanced with KG-RAG offer advantages over larger or commercial LLMs across different task scenarios?
- **RQ3** (Sec. 4.3.3): How effective are different configurations of BOS-LLMs with KG-RAG across different task scenarios?

4.2 Experimental Setup

As discussed in Sec. 3, this paper adopts 7 datasets under different task scenarios to compare 17 raw LLMs and 2 backbone LLMs driven by 6 existing KG-RAG methods (KGRAG, ToG, MindMap, RoK, KGGPT, and Pilot). Qwen1.5-7B and Llama2-7B are employed as the backbone open-source LLMs (BOS-LLMs) to ensure reproducibility and transparency. Note that two KBQA datasets and resutlts are attached in Appx. A.2.

As for the evaluation metrics, we adopt a variety of different metrics. Correct, Wrong, Fail are used for those with ground truth (e.g., CommonsenseQA, CMB-Exam, ExplainCPE), where "Fail" indicates the model fails to generate any answer. As the ExplainCPE also includes explanations, we further use Precision, Recall, F1 to evaluate the quality of generated answers. Besides, we employ BERTScore, ROUGEScore, and G-Eval (Liu et al., 2023) to assess the semantic similarity and overall quality of the answer.

4.3 Main Empirical Analysis

4.3.1 Can KG-RAG improve BOS-LLMs?

In this subsection, we compare the performance of BOS-LLMs with KG-RAG methods to that of BOS-LLMs in Tab. 2, 3, 4, 5, 6, and 13.

Regarding Task Domain. In Tab. 2, 3, 5, 6 and 13, we can observe that KG-RAG methods deliver significant performance improvements across various tasks, including Open-domain QA (CommonsenseQA), Domain-specific QA (GenMedGPT-5K), and Domain-specific Exams (CMB-Exam, ExplainCPE). This demonstrates the effectiveness of KG-RAG in enhancing BOS-LLMs. The only exception is CMCQA in Tab. 4, suggesting that the potential of KG-RAG in clinical scenarios requires further exploration.

Table 5: CMB under Medical Practitioner, Medical Technology, Nursing, and Pharmacy (Self-Construted KG)

		CMB											
		Medical Practitioner			**Medical Technology**			**Nursing**			**Pharmacy**		
Type	**Method**	**Correct**	**Wrong**	**Fail**	**Correct**	**Wrong**	**Fail**	**Correct**	**Wrong**	**Fail**	**Correct**	**Wrong**	**Fail**
LLM only	Qwen1.5-7B	64.06	34.94	1.00	56.71	43.09	0.20	75.55	23.65	0.80	63.93	36.07	0.00
	Qwen2.5-7B	71.74	28.06	0.20	66.93	33.07	0.00	80.96	18.84	0.20	77.56	22.24	0.20
	Qwen2-72B	**84.57**	**15.43**	0.00	**83.97**	15.83	0.20	**89.78**	10.22	0.00	**90.18**	9.82	0.00
	Deepseek-v2l	49.30	49.90	0.80	42.89	55.31	1.80	56.83	42.17	1.00	51.00	47.79	1.20
	ChatGLM-9B	67.54	32.46	0.00	61.92	38.08	0.00	78.31	21.69	0.00	65.66	34.34	0.00
	Yi-34B	72.55	27.45	0.00	67.54	32.46	0.00	83.17	16.83	0.00	78.36	21.44	0.20
	OBLLM-70B	55.71	43.89	0.40	58.72	41.08	0.20	66.27	33.53	0.20	53.82	45.58	0.60
	GPT4o	78.96	20.84	0.20	77.35	22.44	0.20	83.13	16.87	0.00	72.89	26.91	0.20
	o1-mini	65.93	34.07	0.00	71.94	27.86	0.20	74.50	25.50	0.00	60.44	39.56	0.00
	Claude3.5-S	72.34	27.66	0.00	71.74	28.26	0.00	75.90	24.10	0.00	65.86	34.14	0.00
	Gemini1.5-P	74.15	25.85	0.00	70.74	29.26	0.00	80.72	19.28	0.00	70.68	29.32	0.00
KG-RAG (Qwen1.5 -7B)	KGRAG	75.16	23.19	1.66	71.31	27.44	1.25	84.82	14.35	0.83	76.53	21.84	1.63
	ToG	68.74	31.26	0.00	63.73	35.07	1.20	75.75	23.85	0.40	71.14	28.26	0.60
	MindMap	72.55	27.45	0.00	70.54	28.66	0.80	82.57	17.23	0.20	76.75	22.65	0.60
	RoK	74.67	25.33	0.00	71.67	28.33	0.00	85.21	14.79	0.00	<u>77.78</u>	22.22	0.00
	KGGPT	64.40	35.60	0.00	58.63	41.37	0.00	75.60	24.40	0.00	68.40	31.60	0.00
	Pilot	<u>75.35</u>	24.45	0.20	<u>72.95</u>	26.05	1.00	<u>85.37</u>	14.43	0.20	77.35	22.04	0.60

Regarding Task Difficulty: After comparing the performance of BOS-LLMs with KG-RAG in Tab. 2, 3, 5, and 13 with those in Tab. 4 and 6, we can observe that KG-RAG achieve greater improvements in Tab. 2, 3, 5, and 13. BOS-LLMs+KG-RAG even slightly degrade BOS-LLMs in CMCQA. We primarily attribute this to the stronger effectiveness of KG-RAG in lower-difficulty tasks. Compared with CMCQA (Tab. 4) and ExplainCPE (Tab. 6), CommonsenseQA (Tab. 2), GenMedGPT-5K (Tab. 3), and CMB-Exam (Tab. 5 and 13) are relatively easy tasks in each domain because they have a smaller number of L2 questions (see Tab. 1). Thus, the current KG-RAG methods may be able to help BOS-LLMs better utilize external knowledge for easier tasks, but fail to handle hard tasks.

We further delve deeper into this conclusion from KG quality and KBQA tasks. First, the unexpected performance of KG-RAG methods may be caused by the insufficient quality of KGs. In Tab. 6, we replaced the original KG of ExplainCPE (CMCKG) with a specialized self-constructed KG (spKG). The performance using high-quality spKG significantly outperforms that of CMCKG. Second, we exploy KBQA datasets WebQSP (tau Yih et al., 2016) and CWQ (Talmor and Berant, 2018) in Appx. A.2 and reveal that KG-RAG shows outstanding performance on CWQ.

Table 6: ExplainCPE (BOS-LLM is Qwen1.5-7B)

		ExplainCPE			
Type	**Method**	**Correct**	**Wrong**	**Fail**	**F1**
LLM only	Qwen1.5-7B	60.08	39.92	0.00	73.75
	Qwen2.5-7B	69.76	30.24	0.00	75.30
	Qwen2-72B	**81.82**	**18.18**	0.00	**75.75**
	Deepseek-v2l	54.94	45.06	0.00	73.64
	ChatGLM-9B	68.77	31.23	0.00	75.06
	Yi-34B	72.33	27.67	0.00	74.80
	OBLLM-70B	62.85	37.15	0.00	73.07
	GPT4	79.64	20.16	0.20	74.58
	o1-mini	75.10	24.31	0.59	74.19
	Claude3.5-S	76.88	23.12	0.00	75.07
	Gemini1.5-P	69.37	20.75	9.88	67.45
KG-RAG (CMCKG)	KGRAG	58.22	39.60	2.18	74.06
	ToG	<u>61.07</u>	38.74	0.20	74.36
	MindMap	56.92	43.08	0.00	72.01
	RoK	58.29	41.71	0.00	<u>74.99</u>
	KGGPT	53.00	47.00	0.00	74.28
	Pilot	55.93	44.07	0.00	72.53
KG-RAG (spKG)	KGRAG	69.88	29.51	0.61	74.29
	ToG	68.58	31.42	0.00	<u>74.45</u>
	MindMap	70.68	29.32	0.00	72.28
	RoK	<u>74.63</u>	25.37	0.00	74.39
	KGGPT	63.69	36.31	0.00	74.14
	Pilot	73.26	26.74	0.00	73.37

4.3.2 Can BOS-LLMs with KG-RAG are better than commercial LLMs?

In this subsection, we compare the performance of BOS-LLMs with KG-RAG methods to that of commercial LLMs in Tab. 2, 3, 4, 5, 6, and 13.

Regarding Task Domain. For open-domain QA, commercial LLMs significantly outperform BOS-LLMs with KG-RAG methods in CommonsenseQA (Tab. 2), as commercial LLMs may have already internalized sufficient commonsense knowledge. In domain-specific tasks, BOS-LLMs with KG-RAG methods can match or even surpass some commercial LLMs as shown in Tab. 3, 4, 5, 13, and 6. Experimental results show that, given the economic advantages of BOS-LLMs over commercial LLMs, BOS-LLMs enhanced with KG-RAG play a more significant role and remain valuable.

Regarding Task Difficulty. In relatively low-difficulty domain-specific tasks (Tab. 3, 5, and 13), BOS-LLMs with KG-RAG can achieve per-

Table 7: Pre-Retrieval Query enhancement results

Datasets	Methods	Acc	BERT Score			ROUGE Score	
			Precision	Recall	F1	ROUGE-1	ROUGE-L
GenMedGPT-5K	w/o Enhancement	-	0.6499	0.6402	0.6443	**0.2901**	**0.1802**
	Understand (Pilot)	-	**0.6584**	0.6449	**0.6509**	0.2849	0.1785
	Expanse (RoK)	-	0.5941	**0.7110**	0.6468	0.2357	0.1409
	Decompose (KGGPT)	-	0.5687	0.6807	0.6192	0.1850	0.1093
CMCQA	w/o Enhancement	-	0.6660	0.6985	0.6805	0.1370	0.0728
	Understand (Pilot)	-	0.6612	**0.7048**	0.6817	0.1390	0.0733
	Expanse (RoK)	-	0.6619	0.6973	0.6785	**0.1529**	**0.0800**
	Decompose (KGGPT)	-	**0.6677**	0.7040	**0.6848**	0.1513	0.0787
ExplainCPE	w/o Enhancement	66.80	0.7279	0.7537	0.7354	0.3020	0.1963
	Understand (Pilot)	73.26	**0.7281**	0.7515	0.7337	0.3000	0.1950
	Expanse (RoK)	**74.63**	0.7242	**0.7670**	**0.7439**	0.2961	0.1973
	Decompose (KGGPT)	63.69	0.7223	0.7638	0.7414	**0.3169**	**0.2103**

Table 8: Configurations comparison on GenMedGPT-5K

Config	BERT Score			ROUGE Score			G-Eval			
	Prec.	Rec.	F1	R-1	R-2	R-L	CR	Comp	Corr	Emp
Facts_w/o Prompt	56.62	67.74	61.64	17.16	3.44	10.31	99.95	97.83	99.24	79.29
Facts+CoT	58.51	59.53	59.00	25.32	4.91	14.70	99.77	87.44	97.36	78.22
Facts+ToT	64.42	63.43	63.91	28.54	5.86	17.53	79.30	62.11	69.86	55.07
Facts+MindMap	65.10	63.78	64.37	**28.70**	6.02	17.82	99.49	82.10	96.62	77.64
Path_w/o Prompt	56.85	**68.00**	61.89	18.32	3.77	10.89	**100.00**	**97.93**	**99.29**	**79.44**
Path+CoT	58.02	59.01	58.50	25.05	4.82	14.53	99.91	87.09	98.83	79.28
Path+ToT	63.85	62.92	63.41	28.42	5.75	17.43	83.00	65.60	77.00	60.00
Path+MindMap	**65.84**	64.49	**65.09**	28.49	**6.07**	**17.85**	99.54	81.33	97.56	78.10
Subgraph_w/o Prompt	56.40	67.01	61.21	16.84	3.26	10.30	98.43	94.39	97.93	78.08
Subgraph+CoT	58.49	61.43	59.85	25.21	4.72	14.47	99.47	88.67	96.76	77.77
Subgraph+ToT	57.83	59.92	58.94	25.38	4.96	14.92	75.32	57.91	65.27	51.24
Subgraph+MindMap	59.29	58.01	58.60	26.16	5.57	16.13	97.13	79.21	92.42	74.44

formance comparable to or even surpass that of commercial LLMs. This suggests that KG-RAG effectively mitigates the knowledge limitations of BOS-LLMs, Enable them to be competitive in easier tasks. However, in Tab. 4 and 6 with more L2 questions, BOS-LLMs still lag behind commercial LLMs overall, even if KG-RAGs are able to narrow the performance gap. In hard tasks, commercial LLMs likely benefit not only from their extensive knowledge but also from stronger reasoning and generalization abilities, which could further inspire the future development of KG-RAG.

4.3.3 How effective are different KG-RAG configurations?

In this subsection, we compare the performance of diffrent KG-RAG configurations in Tab. 7, 8, 9, 10 on GenMedGPT-5K, CMCQA, ExplainCPE.

Impact of Query Enhancement. In Tab. 7, we compare the impact of different query enhancement methods, including query understanding (Pilot), query expansion (RoK), and query decomposition (KGGPT). Given no single method shows absolute superiority, we may analyze the reustls from the perspective of the length of questions.

For datasets with shorter question lengths (GenMedGPT-5K, ExplainCPE): understanding and expansion methods are relatively effective, while decomposition one performs poorly, possibly because single-sentence questions do not require further decomposition. For longer medical dialogue questions (CMCQA), decomposition appears to be slightly advantageous with the highest F1 score. Overall, query understanding shows robustness, but with limited improvement effects. Query expansion may be more suitable for short questions, while query decomposition may be more suitable for long questions.

Impact of Retrieval Forms. In Tab. 8, 9, and 10, we compare the impact of different retrieval forms in KG-RAG, including fact, path, and subgraph.

On GenMedGPT-5K (Tab. 8), using facts and paths as retrieval forms typically outperforms subgraphs in terms of BERT and ROUGE Scores. Similarly, using facts as retrieval forms shows better performance on ExplainCPE (Tab. 10). This suggests that for short questions, providing retrieval forms of fact or path might be more conducive to generating answers with better semantic similarity, while subgraphs might introduce redundant noises. As for G-Eval metrics, the differences between various retrieval forms are minor. This suggests

Table 9: Configurations comparison on CMCQA

Config	BERT Score			ROUGE Score			G-Eval			
	Prec.	Rec.	F1	R-1	R-2	R-L	CR	Comp	Corr	Emp
facts_w/o Prompt	66.05	**70.48**	68.12	**14.00**	**1.33**	7.39	**100.0**	**100.0**	**100.0**	**100.0**
facts+CoT	65.46	69.37	67.30	13.43	1.05	7.63	98.69	95.08	97.70	96.72
facts+ToT	64.13	68.80	66.32	12.71	0.94	7.40	97.70	94.10	96.39	96.07
facts+MindMap	64.20	68.17	66.11	12.49	0.95	7.29	96.07	92.13	93.77	92.79
path_w/o Prompt	**66.12**	**70.48**	**68.17**	13.90	1.22	7.33	**100.0**	**100.0**	99.67	**100.0**
path+CoT	65.40	69.46	67.30	13.14	1.10	7.40	96.07	90.49	93.44	93.77
path+ToT	64.16	68.89	66.38	12.57	0.99	7.22	97.38	92.79	95.74	93.77
path+MindMap	64.13	68.06	65.98	12.33	0.95	7.31	92.79	87.87	89.84	89.18
Subgraph_w/o Prompt	66.11	70.45	68.15	13.91	1.31	7.35	99.34	99.34	99.02	99.34
Subgraph+CoT	65.42	69.62	67.39	13.90	1.18	**7.82**	96.39	94.75	95.08	95.74
Subgraph+ToT	64.12	68.83	66.33	12.71	1.06	7.35	98.03	95.74	97.05	96.39
Subgraph+MindMap	64.17	67.96	65.96	12.45	0.97	7.25	91.48	87.54	90.16	88.85

Table 10: Configurations comparison on ExplainCPE

Config	Acc	BERT Score			ROUGE Score			G-Eval			
		Prec.	Rec.	F1	R-1	R-2	R-L	CR	Comp	Corr	Emp
Facts_w/o Prompt	**73.26**	72.81	75.15	73.37	**30.00**	**9.47**	**19.50**	**95.83**	**90.87**	**92.27**	**86.28**
Facts+CoT	69.83	71.20	**78.33**	**74.52**	22.35	7.30	14.74	79.80	79.94	79.74	80.30
Facts+ToT	65.91	68.92	77.71	72.98	16.56	5.31	10.89	79.50	79.86	80.06	79.94
Facts+MindMap	59.50	67.12	75.22	70.85	15.43	5.09	10.30	79.59	79.51	80.01	79.53
Path_w/o Prompt	63.22	72.96	69.83	69.43	26.50	8.11	17.21	94.02	84.45	91.09	82.68
Path+CoT	58.68	**76.11**	77.46	74.12	21.14	7.00	14.40	79.75	79.89	79.95	79.94
Path+ToT	56.20	**76.11**	77.10	73.89	20.57	6.99	14.22	79.82	79.59	79.75	79.83
Path+MindMap	55.37	67.07	75.24	70.84	14.97	4.80	9.97	80.08	80.04	80.34	79.64
Subgraph_w/o Prompt	66.74	71.06	63.91	65.14	15.62	6.01	11.96	94.97	84.00	91.79	82.56
Subgraph+CoT	63.22	71.06	78.32	74.44	21.87	7.11	14.47	80.30	80.27	80.06	80.53
Subgraph+ToT	61.16	68.89	77.70	72.97	16.46	5.27	10.88	80.06	79.72	79.93	80.16
Subgraph+MindMap	56.20	67.14	75.18	70.84	15.36	4.96	10.21	80.50	80.38	79.79	80.23

that G-Eval, as a LLM-based measurement, might be influenced by the quality of answers rather than subtle differences in retrieval forms. Different from Tab. 8 and 10, different retrieval forms perform very similarly on CMCQA (Tab 9). This indicates that for the long dialogue diagnosis task retrieval form may not be a key factor affecting performance.

Impact of Prompt Strategy. In Tab. 8, 9, and 10, we compare the impact of prompt strategies in KG-RAG, including CoT, ToT, and MindMap.

On GenMedGPT-5K (Tab. 8), w/o prompt significantly outperforms prompt strategies in G-Eval metrics. However, prompts strategies (especially MindMap) perform better in BERT and ROUGE Scores compared to w/o prompt. Similar observations are also found in ExplainCPE (Tab. 10), where removing prompt strategies significantly outperforms strategies using prompts in Acc and G-Eval metrics. These observations suggest that for domain-specific tasks like GenMedGPT-5K and ExplainCPE, not using prompt strategies still better aligns with overall answer quality assessment (Acc & G-Eval), while using prompts might improve language quality but at the cost of overall answer quality. On CMCQA (Tab. 9), removing prompt strategy significantly outperforms strategies using prompts across all metrics (BERT, ROUGE, G-Eval). This indicates that for long dialogue diagnosis, prompt strategies not only provide no benefit but actually degrade performance.

5 Conclusion

This study systematically explores the applicability and configuration strategies of KG-RAG across different task scenarios. Experimental results indicate that KG-RAG can significantly enhance the performance of BOS-LLMs in domain-specific tasks. However, KG-RAG's benefits are relatively limited in open domains. Furthermore, we observe that as difficulty increases, improvement magnitude becomes constrained. Through detailed analysis of KG-RAG configurations, we find that there is no universally optimal query enhancement method, with the best strategy depending on task properties. The retrieval forms do not have a deterministic impact on performance, though path and facts may hold slight advantages. Notably, in domain-specific tasks, removing prompts typically performs best on G-Eval metric, suggests that generating answers directly from retrieved knowledge may better meet practical requirements.

6 Limitations

This study primarily focuses on the small-scale LLMs, future works could explore the performance of larger-scale LLMs within KG-RAG methods. KG-RAG's configurable space is vast, future works could delve deeper into exploring KG-RAG configurations across more dimensions. This study preliminarily examines the impact of KG quality on ExplainCPE dataset. Future works could do a more systematic investigation of the quantitative impact of KG quality on KG-RAG performance.

References

Malaikannan Sankarasubbu Ankit Pal. 2024. Openbiollms: Advancing open-source large language models for healthcare and life sciences. https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Banghao Chen, Zhaofeng Zhang, Nicolas Langren'e, and Shengxin Zhu. 2023. Unleashing the potential of prompt engineering in large language models: a comprehensive review. *ArXiv*, abs/2310.14735.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

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*.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *ArXiv*, abs/2404.16130.

Wenqi Fan, Yujuan Ding, Liangbo 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 *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 6491–6501, New York, NY, USA. Association for Computing Machinery.

Chunjing Gan, Dan Yang, Binbin Hu, Hanxiao Zhang, Siyuan Li, Ziqi Liu, Yue Shen, Lin Ju, Zhiqiang Zhang, Jinjie Gu, Lei Liang, and Jun Zhou. 2024. Similarity is not all you need: Endowing retrieval augmented generation with multi layered thoughts. *ArXiv*, abs/2405.19893.

Michael Glass, Gaetano Rossiello, Md Faisal Mahbub Chowdhury, Ankita Naik, Pengshan Cai, and Alfio Gliozzo. 2022. Re2G: Retrieve, rerank, generate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2701–2715, Seattle, United States. Association for Computational Linguistics.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Xinyan Guan, Yanjiang Liu, Hongyu Lin, Yaojie Lu, Ben He, Xianpei Han, and Le Sun. 2024. Mitigating large language model hallucinations via autonomous knowledge graph-based retrofitting. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024a. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Xinke Jiang, Ruizhe Zhang, Yongxin Xu, Rihong Qiu, Yue Fang, Zhiyuan Wang, Jinyi Tang, Hongxin Ding, Xu Chu, Junfeng Zhao, and Yasha Wang. 2024b. Hykge: A hypothesis knowledge graph enhanced framework for accurate and reliable medical llms responses. *Preprint*, arXiv:2312.15883.

Jiho Kim, Yeonsu Kwon, Yohan Jo, and Edward Choi. 2023. KG-GPT: A general framework for reasoning on knowledge graphs using large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9410–9421, Singapore. Association for Computational Linguistics.

Dongfang Li, Jindi Yu, Baotian Hu, Zhenran Xu, and Min Zhang. 2023a. ExplainCPE: A free-text explanation benchmark of Chinese pharmacist examination. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1922–1940, Singapore. Association for Computational Linguistics.

Mufei Li, Siqi Miao, and Pan Li. 2025a. Simple is effective: The roles of graphs and large language models in knowledge-graph-based retrieval-augmented generation. In *International Conference on Learning Representations*.

Siran Li, Linus Stenzel, Carsten Eickhoff, and Seyed Ali Bahrainian. 2025b. Enhancing retrieval-augmented generation: A study of best practices. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6705–6717, Abu Dhabi, UAE. Association for Computational Linguistics.

Xiaonan Li, Changtai Zhu, Linyang Li, Zhangyue Yin, Tianxiang Sun, and Xipeng Qiu. 2024. LLatrieval: LLM-verified retrieval for verifiable generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5453–5471, Mexico City, Mexico. Association for Computational Linguistics.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023b. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6).

Zhuoqun Li, Xuanang Chen, Haiyang Yu, Hongyu Lin, Yaojie Lu, Qiaoyu Tang, Fei Huang, Xianpei Han, Le Sun, and Yongbin Li. 2025c. StructRAG: Boosting knowledge intensive reasoning of LLMs via inference-time hybrid information structurization. In *The Thirteenth International Conference on Learning Representations*.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3580–3599.

Boci Peng, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, and Siliang Tang. 2024. Graph retrieval-augmented generation: A survey. *Preprint*, arXiv:2408.08921.

Pranab Sahoo, Prabhash Meharia, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024a. A comprehensive survey of hallucination in large language, image, video and audio foundation models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11709–11724, Miami, Florida, USA. Association for Computational Linguistics.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Sohel Mondal, and Aman Chadha. 2024b. A systematic survey of prompt engineering in large language models: Techniques and applications. *ArXiv*, abs/2402.07927.

Ahmmad O. M. Saleh, Gokhan Tur, and Yucel Saygin. 2024. SG-RAG: Multi-hop question answering with large language models through knowledge graphs. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, pages 439–448, Trento. Association for Computational Linguistics.

Zhihong Shao, Damai Dai, Daya Guo, Bo Liu (Benjamin Liu), Zihan Wang, and Huajian Xin. 2024. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *ArXiv*, abs/2405.04434.

Karthik Soman, Peter W Rose, John H Morris, Rabia E Akbas, Brett Smith, Braian Peetoom, Catalina Villouta-Reyes, Gabriel Cerono, Yongmei Shi, Angela Rizk-Jackson, Sharat Israni, Charlotte A. Nelson, Sui Huang, and Sergio Baranzini. 2023. Biomedical knowledge graph-optimized prompt generation for large language models. *Bioinformatics*, 40.

Yuan Sui and Bryan Hooi. 2024. Can knowledge graphs make large language models more trustworthy? an empirical study over open-ended question answering. *ArXiv*, abs/2410.08085.

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Sai Wang, Chen Lin, Yeyun Gong, Lionel M. Ni, Heung yeung Shum, and Jian Guo. 2023. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. In *International Conference on Learning Representations*.

Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* *Volume 1 (Long Papers)*, pages 641–651, New Orleans, Louisiana. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Wen tau Yih, Matthew Richardson, Christopher Meek, Ming-Wei Chang, and Jina Suh. 2016. The value of semantic parse labeling for knowledge base question answering. In *Annual Meeting of the Association for Computational Linguistics*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Qwen Team. 2024. Introducing qwen1.5.

S. M Towhidul Islam Tonmoy, S M Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *ArXiv*, abs/2401.01313.

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*.

Song Wang, Junhong Lin, Xiaojie Guo, Julian Shun, Jundong Li, and Yada Zhu. 2025. Reasoning of large language models over knowledge graphs with super-relations. In *The Thirteenth International Conference on Learning Representations*.

Xidong Wang, Guiming Chen, Song Dingjie, Zhang Zhiyi, Zhihong Chen, Qingying Xiao, Junying Chen, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, and Haizhou Li. 2024a. CMB: A comprehensive medical benchmark in Chinese. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6184–6205, Mexico City, Mexico. Association for Computational Linguistics.

Yuqi Wang, Boran Jiang, Yi Luo, Dawei He, Peng Cheng, and Liangcai Gao. 2024b. Reasoning on efficient knowledge paths: Knowledge graph guides large language model for domain question answering. *arXiv preprint arXiv:2404.10384*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022b. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Yilin Wen, Zifeng Wang, and Jimeng Sun. 2024. MindMap: Knowledge graph prompting sparks graph of thoughts in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10370–10388, Bangkok, Thailand. Association for Computational Linguistics.

Fei Xia, Bin Li, Yixuan Weng, Shizhu He, Kang Liu, Bin Sun, Shutao Li, and Jun Zhao. 2022. MedConQA: Medical conversational question answering system based on knowledge graphs. In *Proceedings of the The 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 148–158, Abu Dhabi, UAE. Association for Computational Linguistics.

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*.

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

Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze Daniel Gui, Ziran Will Jiang, Ziyu Jiang, Lingkun Kong, Brian Moran, Jiaqi Wang, Yifan Ethan Xu, An Yan, Chenyu Yang, Eting Yuan, Hanwen Zha, Nan Tang, Lei Chen, Nicolas Scheffer, Yue Liu, Nirav Shah, Rakesh Wanga, Anuj Kumar, Wen tau Yih, and Xin Luna Dong. 2024c. Crag – comprehensive rag benchmark. *arXiv preprint arXiv:2406.04744*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Miao Zhang, Rufeng Dai, Ming Dong, and Tingting He. 2022. DRLK: Dynamic hierarchical reasoning with language model and knowledge graph for question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5123–5133, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Qinggang Zhang, Shengyuan Chen, Yuan-Qi Bei, Zheng Yuan, Huachi Zhou, Zijin Hong, Junnan Dong, Hao Chen, Yi Chang, and Xiao Huang. 2025. A survey of graph retrieval-augmented generation for customized large language models.

Siyun Zhao, Yuqing Yang, Zilong Wang, Zhiyuan He, Luna K. Qiu, and Lili Qiu. 2024. Retrieval augmented generation (rag) and beyond: A comprehensive survey on how to make your llms use external data more wisely. *ArXiv*, abs/2409.14924.

Lexin Zhou, Wout Schellaert, Fernando Martínez-Plumed, Yael Moros-Daval, Cèsar Ferri, and José Hernández-Orallo. 2024. Larger and more instructable language models become less reliable. *Nature*, pages 1–8.

A Complementary Experiments

A.1 Datasets

The detailed descriptions of the adopted KG-RAG datasets are summarized as follows:

- CommonsenseQA (Talmor et al., 2019) is a multiple-choice QA dataset specifically designed to evaluate commonsense reasoning capabilities. Each question is accompanied by five candidate answers, only one of which is correct.
- GenMedGPT-5K (Li et al., 2023b) is a medical dialogue dataset, covers patient-doctor single round dialogues. Generated through interactions between GPT-3.5 and the iCliniq disease database, this dataset contains clinical conversations covering patient symptoms, diagnoses, recommended treatments, and diagnostic tests.
- CMCQA (Xia et al., 2022) is a comprehensive medical conversational QA dataset derived from professional Chinese medical consultation platform. The dataset encompasses multi-round clinical dialogues across 45 medical specialties, including andrology, stomatology, and obstetrics-gynecology, representing diverse clinical interactions between healthcare providers and patients.
- CMB-Exam (Wang et al., 2024a) covers 280,839 questions from six major medical professional qualification examinations, including physicians, nurses, medical technologists and pharmacists, as well as Undergraduate Disciplines Examinations and Graduate Entrance Examination in the medical field at China. Given the extensive scale of CMB-Exam, we sample a subset of CMB-Exam that comprise 3,000 questions, where 500 questions are randomly sampled from each category.
- ExplainCPE (Li et al., 2023a) is a Chinese medical benchmark dataset containing over 7K instances from the National Licensed Pharmacist Examination. This dataset is distinctive in providing both multiple-choice answers and their corresponding explanations.

Additionally, we incorporated two representative KBQA datasets, WebQSP (tau Yih et al., 2016) and Complex Web Questions (CWQ) (Talmor and Berant, 2018), discussing KBQA as a special case. WebQSP consists of natural language questions emphasizing single-hop factoid queries, while CWQ features more complex multi-hop questions requiring compositional reasoning over knowledge graphs.

A.2 KBQA experimental results

We conducted experiments on two KBQA datasets and the results are shown in Tab. 11.

Table 11: Experimental Results on KBQA Datasets

Type	Method	WebQSP	CWQ
KG-RAG	MindMap (Llama2-7B)	30.82	30.51
	ToG (ChatGPT) (Sun et al., 2023)	**75.80**	**58.90**
LLM only	Llama3.2-1B	37.45	15.21
	Llama2-7B	43.29	21.91
	Llama3-8B	55.41	28.09
	Llama2-70B	53.68	28.87
	Mixtral-8*7B	58.01	33.25
	ChatGPT	63.30	37.60

A.3 The remaining results of CMB-Exam

Due to space constraints, the remaining experimental results of the CMB-Exam are shown in Tab 13.

A.4 Other experimental settings

Our KG-RAG framework is built on LangChain[1]. The local open-source LLMs are deployed based on the llama.cpp[2] project. Except for the context window size, which is adjusted according to the dataset, all other parameters use default configurations, such as temperature is 0.8. Both LangChain and llama.cpp are open-source projects, providing good transparency and reproducibility.

[1]https://www.langchain.com/
[2]https://github.com/ggml-org/llama.cpp

Table 12: Prompt Example for Knowledge Graph Construction

```
prompt = f"""As a professional knowledge extraction assistant, your task is to extract knowledge triples from the given question.
1. Carefully read the question description, all options, and the correct answer.
2. Focus on the core concept "{question_concept}" in the question.
3. Extract commonsense knowledge triples related to the question.
4. Each triple should be in the format: subject\tpredicate\tobject
5. Focus on the following types of relationships:
 - Conceptual relations
 - Object properties
 - Object functions
 - Spatial relations
 - Temporal relations
 - Causal relations
6. Each triple must be concrete and valuable commonsense knowledge.
7. Avoid subjective or controversial knowledge.
8. Ensure triples are logically sound and align with common sense.
Please extract knowledge triples from this multiple-choice question:
Question: {question}
Core Concept: {question_concept}
Correct Answer: {correct_answer}
Please output knowledge triples directly, one per line, in the format: subject\tpredicate\tobject. """
```

For the evaluation, we employed Bert Score metrics using "bert-base-uncased (Devlin et al., 2018)" and "bert-base-chinese[3]" models to evaluate English and Chinese results respectively, while ROUGE Score version 0.1.2 was utilized. Due to resource constraints, G-Eval assessments were conducted using locally deployed LLMs, with Llama2-70B for English tasks and Qwen2-72B for Chinese tasks.

B Knowledge Graph Construction

Apart from EMCKG for GenMedGPT-5K and CMCKG for CMCQA (Wen et al., 2024), we employed a consistent KG construction method for other datasets, utilizing LLMs to extract knowledge triples from the datasets to build specialized KGs. The prompt example is shown in Tab. 12.

For the KBQA task, we referenced ToG (Sun et al., 2023) and deployed Freebase using the Virtuoso[4] graph database. All other KGs used in the datasets were deployed using Neo4j[5].

C Ethics Statement

We are committed to responsible AI development by focusing on improving model accuracy through knowledge graph integration while maintaining transparency in our research methodology. This research utilized publicly available datasets for experimental evaluation of KG-RAG. While these datasets are commonly used benchmarks, we acknowledge their potential inherent biases, particularly in medical domain datasets where healthcare disparities and demographic representation must be considered. Our study aims to improve KG-RAG methodologies for academic purposes, and we emphasize that any real-world applications, especially in healthcare, would require additional validation and ethical review.

Table 13: CMB under Postgraduate and Professional (Self-Constructed KG), BOS-LLM is Qwen1.5-7B

		CMB					
		Postgraduate			Professional		
Type	**Method**	**Correct**	**Wrong**	**Fail**	**Correct**	**Wrong**	**Fail**
LLM only	Qwen1.5-7B	61.04	36.14	2.81	64.13	35.07	0.80
	Qwen2.5-7B	80.36	19.64	0.00	74.15	25.85	0.00
	Qwen2-72B	**87.78**	12.02	0.20	**83.97**	16.03	0.00
	Deepseek-v2l	52.71	45.29	2.00	49.70	47.70	2.61
	ChatGLM-9B	71.74	28.26	0.00	68.94	31.06	0.00
	Yi-34B	75.55	24.45	0.00	74.75	25.25	0.00
	OBLLM-70B	60.32	37.88	1.80	64.73	34.47	0.80
	GPT4o	76.95	22.44	0.60	78.96	21.04	0.00
	o1-mini	63.13	36.27	0.60	73.55	26.45	0.00
	Claude3.5-S	69.54	30.46	0.00	73.75	26.25	0.00
	Gemini1.5-P	75.95	24.05	0.00	77.56	22.44	0.00
KG-RAG	KG-RAG	76.05	21.64	2.31	74.84	23.04	2.11
	ToG	72.75	27.05	0.20	67.74	32.06	0.20
	MindMap	78.11	21.49	0.40	74.55	25.05	0.40
	RoK	72.73	27.27	0.00	73.63	26.27	0.00
	KGGPT	70.99	29.01	0.00	66.33	33.67	0.00
	Pilot	80.16	19.44	0.40	76.91	23.09	0.00

[3] https://huggingface.co/google-bert/bert-base-chinese
[4] http://virtuoso.openlinksw.com/
[5] https://neo4j.com/