HyP-KGRAG: Hypothetical Path-Based Knowledge Graph Retrieval Augmented Generation with DeepSeek

Zhaotai Liu[1], Harald Sack[1,2] and Genet Asefa Gesese[1,2]

[1]Karlsruhe Institute of Technology, Institute AIFB, Germany
[2]FIZ Karlsruhe, Germany

Abstract

Large Language Models (LLMs) often produce hallucinated or factually incorrect responses in domain-specific question answering (QA) tasks. To address this limitation, this work explores the integration of Knowledge Graphs (KGs) with Retrieval-Augmented Generation (RAG) as a strategy to improve factual accuracy and multi-hop evidence selection. Specifically, the benefits of using structured information from a KG to enhance the QA performance of the DeepSeek model are investigated. A novel framework, HyP-KGRAG, is introduced in which KG triples are retrieved via hypothetical paths and refined through an LLM-based denoising module. Experimental results on the material science MSE-KG dataset show that HyP-KGRAG significantly improves the QA performance of DeepSeek and other baseline models, achieving a ROUGE-1 F1 score of 0.532 and an SBERT similarity of 0.629.

Keywords

Knowledge Graph, Large Language Models, Question Answering, Retrieval Augmented Generation

1. Introduction

Large language models (LLMs) have attracted significant attention due to their remarkable capabilities in various natural language processing (NLP) tasks, including question answering (QA) [1], text summarization [2], and machine translation [3], driven by massive data, parameters, and computational resources. However, their parametric knowledge is limited by training data timeliness and domain coverage, leading to hallucinations, i.e., factually incorrect or ungrounded outputs. Traditional solutions to mitigate hallucinations typically rely on retraining or fine-tuning LLMs with updated data, a process that is computationally costly and inflexible. An alternative and more efficient approach is Retrieval-Augmented Generation (RAG) [4], which dynamically incorporates external knowledge sources, such as documents or knowledge graphs (KGs), into the generation process. RAG improves factual accuracy and relevance without necessitating costly retraining, thereby overcoming some limitations of standalone LLMs. KGs, with their structured representation of information as triples, comprising a head entity, a relation, and a tail entity, play a vital role in enhancing RAG systems. They provide a rich, explicit, and semantically organized knowledge, which can be leveraged to improve reasoning, factual consistency, and explainability in QA tasks.

This work focuses on addressing the following research question: *How can KGs be effectively leveraged within a RAG framework to reduce hallucinations and enhance domain-specific QA performance in LLMs such as DeepSeek?*

The key contributions of this work are as follows:

- **HyP-KGRAG framework**: a novel RAG-based framework that retrieves KG triples via hypothetical paths and leverages LLMs for denoising and answer generation.
- **Experiments on a materials science KG**: demonstrate that HyP-KGRAG significantly enhances domain-specific QA performance in terms of accuracy and semantic coherence.

RAGE-KG 2025: The Second International Workshop on Retrieval-Augmented Generation Enabled by Knowledge Graphs, co-located with ISWC 2025, November 2–6, 2025, Nara, Japan

uvyrq@student.kit.edu (Z. Liu); harald.sack@fiz-karlsruhe.de (H. Sack); genet-asefa.gesese@fiz-karlsruhe.de (G. A. Gesese)
0009-0004-4740-5790 (Z. Liu); 0000-0001-7069-9804 (H. Sack); 0000-0003-3807-7145 (G. A. Gesese)

- **Ablation studies**: validate the benefits of triple-structured retrieval and the denoising module to achieve a better balance between precision and recall.

2. Preliminaries and Related Work

2.1. Preliminaries

DeepSeek-V3 DeepSeek-V3 [5] is a Mixture-of-Experts (MoE) language model with 671B total parameters, designed with Multi-head Latent Attention and DeepSeekMoE architectures for efficient training and inference, achieving performance competitive with leading closed-source models while maintaining a modest compute footprint.

RAG RAG [4] is a framework that integrates parametric memory, typically a pre-trained sequence-to-sequence language model, with non-parametric memory in the form of a dense index over external knowledge sources such as Wikipedia. This setup allows the model to retrieve relevant information dynamically during generation, improving performance on knowledge-intensive tasks. RAG enhances factual accuracy, diversity, and specificity of responses compared to models relying solely on internal parametric knowledge.

BM25 BM25 (Best Matching 25) [6] is a probabilistic information retrieval model that improves upon traditional TF-IDF by introducing term frequency saturation and document length normalization. It is widely used in modern search engines (e.g., Elasticsearch) due to its robust performance in ranking documents based on query relevance.

BGE-M3 BGE-M3 [7], developed by the Beijing Academy of Artificial Intelligence (BAAI) and released in early 2024, represents a state-of-the-art universal text embedding model with 1024-dimensional representations, capable of cross-task and cross-lingual semantic understanding. As a flagship model in the BGE series, it achieves groundbreaking advancements across three key dimensions through its high-dimensional embedding space.

2.2. Related Work

The integration of KGs into RAG has emerged as a promising direction to enhance the problem of hallucination with LLMs. Several recent works have proposed methods for incorporating structured knowledge into the RAG pipeline, moving beyond traditional text-based retrieval. KAPING [8] introduces a zero-shot KG-RAG framework that retrieves and linearizes relevant KG triples into natural language prompts without requiring model fine-tuning. While this method improves zero-shot performance, it does not leverage the rich structural and multi-hop relationships inherent in KGs. To address this, CommunityKG-RAG [9] utilizes community detection via the Louvain algorithm and aligns KG clusters with queries using BERT embeddings. This structure-aware approach significantly improves fact-checking and contextual relevance in zero-shot settings.

For more complex query understanding and summarization tasks, GraphRAG [10] constructs KGs from input documents and performs topic-oriented clustering using the Leiden algorithm. Although effective for global summarization, its performance on traditional QA tasks remains limited. LightRAG [11] enhances retrieval efficiency and adaptability through graph-structured text indexing and a two-layer retrieval process, enabling improved comprehension of both entity-level and topic-level semantics. Reasoning over KGs has also gained attention. RoG [12] presents a ”planning-retrieval-reasoning” framework that explicitly models reasoning paths over KGs to guide LLMs, resulting in more explainable outputs. In a related effort, GNN-RAG [13] integrates Graph Neural Networks into the RAG pipeline to perform multi-hop reasoning over dense subgraphs, improving answer accuracy in complex KGQA scenarios.

KG2RAG [14] further extends retrieval by combining semantic similarity-based chunk selection with m-hop KG traversal. A graph-structured context organization module filters and arranges retrieved chunks into coherent passages for LLM input, enhancing both coverage and relevance. In professional domains, KAG [15] unifies LLMs and KGs via a hybrid reasoning engine and mutual indexing, enabling logical reasoning and high semantic fidelity in settings like government and medical QA. KGMistral [16] proposes a hybrid framework that augments the Mistral 7B model with rule-based SPARQL queries. It supports one-hop and two-hop question patterns by aligning predefined templates with the logic of user queries. Although effective in scientific domains, KGMistral's reliance on static templates may limit its flexibility across diverse question types.

In contrast to existing methods, the proposed HyP-KGRAG framework introduces a novel hypothetical path-based retrieval approach that leverages structured knowledge from KGs and incorporates LLM-based denoising to enhance QA performance. While it is inspired by HyDE [17], which generates hypothetical documents using instruction-following LLMs such as InstructGPT, HyDE does not utilize KGs. HyP-KGRAG uniquely grounds its hypothetical paths in KG triples, enabling more semantically aligned and factually grounded multi-hop evidence retrieval for domain-specific QA tasks. Moreover, while most prior KG-augmented RAG approaches utilize various LLMs such as GPT and Mistral, this work specifically focuses on integrating KG-based retrieval with the DeepSeek-V3 model. Unlike general-purpose RAG systems, HyP-KGRAG is tailored to address hallucination and domain adaptation challenges observed in DeepSeek during domain-specific QA.

3. HyP-KGRAG

This section introduces HyP-KGRAG (Hypothetical Path-Based KG Retrieval Augmented Generation with DeepSeek), a novel framework designed to enhance the retrieval quality and multi-hop evidence selection of LLMs, with a focus on DeepSeek without loss of generality, through KG integration. As illustrated in Figure 1, HyP-KGRAG contains three main components: offline processing, hypothetical path construction, and LLM-as-Judge denoising and generation.

3.1. Offline Processing

In the offline processing phase, HyP-KGRAG performs two key preparatory tasks: KG verbalization and index construction. These steps transform structured knowledge into a format that enables fast and semantically meaningful retrieval during inference.

KG Verbalization. This step involves converting structured triples (i.e., subject, predicate, object) into natural language documents. Unlike prior approaches that employ LLMs to generate free-form verbalization of triples, HyP-KGRAG uses a linear verbalization strategy, as described in [8]. In this method, each triple is preserved in its original structure by directly concatenating the subject, predicate, and object. This technique preserves the explicit relationships in the KG while minimizing noise from generative paraphrasing. As confirmed by our ablation study in Section 4.5, such linear verbalization consistently outperforms LLM-generated paraphrases in downstream QA tasks.

Index Construction. Following verbalization, the triples are encoded and indexed for retrieval. For semantic retrieval, the bge-M3 embedding model is employed to encode verbalized triples into dense vector representations, supporting multi-granularity encoding. The resulting embeddings are indexed using FAISS [18] to enable efficient approximate nearest neighbor (ANN) search based on inner product similarity during inference.

3.2. Hypothetical Paths Construction

As depicted in Figure 1, HyP-KGRAG leverages an LLM (specifically DeepSeek-V3) to identify potential paths underlying the query, where these hypothetical paths are structured as triplets. These triples

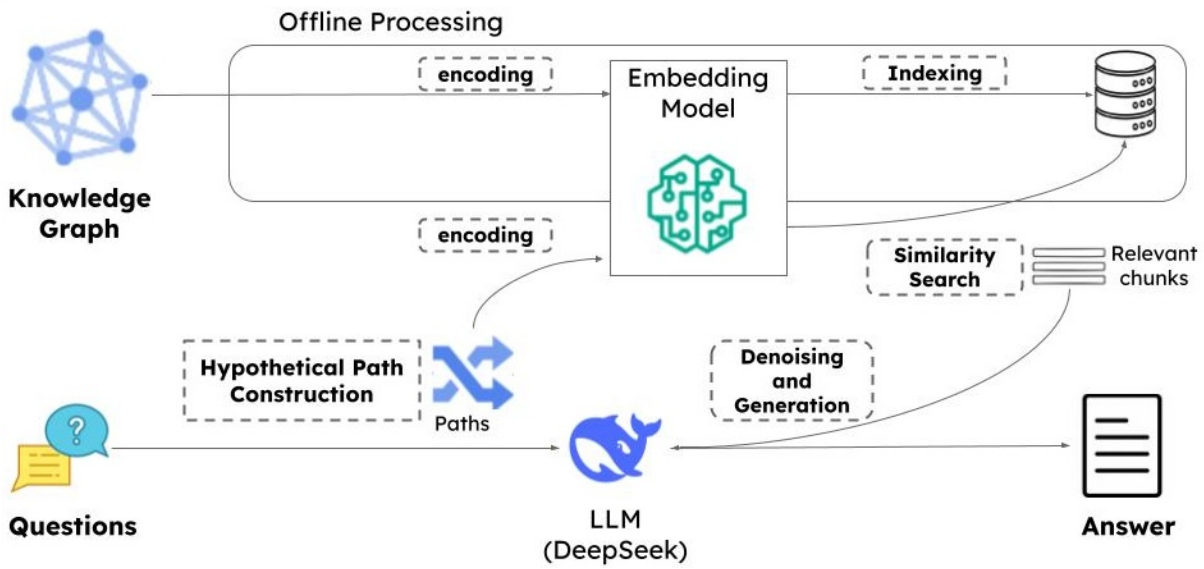

Figure 1: The general architecture of the proposed HyP-KGRAG approach

undergo the same verbalization process as described in Section 3.1. Subsequently, BGE-M3 is employed to encode the text into vector representations. FAISS is then utilized to retrieve semantically closest factual triplets from an offline-constructed vector database, with similarity measured via the inner product of vectors. To construct a complete retrieval paths for the query, HyP-KGRAG utilizes class information to fill in missing components, whether they are target answers or intermediate nodes, ensuring coherence and completeness of the paths. A class represents an abstract collection of entities (e.g., Person, Researcher, Organization), making it well-suited as a bridge in incomplete paths.

Effective multi-hop evidence retrieval over KGs requires generating coherent paths that span one or more hops. Accordingly, user questions are categorized into two types:

Single-hop Questions. These involve a direct relation between two entities, requiring only one retrieval step to obtain the supporting evidence for the answer. These questions are relatively straightforward, as they rely on a single factual assertion represented as a triple in the KG. For example, the question *"Who published the dataset 'Nano2023'?"* requires identifying an entity (likely a person or organization) directly connected to the dataset via the `published` relation. Although single-hop, HyP-KGRAG still constructs hypothetical paths in both directions to support more robust retrieval. Class-constrained hypothetical paths include:

$$\text{Person} \xrightarrow{\text{published}} \text{Nano2023} \quad \text{or} \quad \text{Nano2023} \xrightarrow{\text{published by}} \text{Person}$$

Multi-hop Questions. These involve more complex chains of entities and relations, requiring multi-hop evidence retrieval to gather the information needed to answer the question. For instance, the question *"Which researcher from MIT published the dataset 'Nano2023' cited in 'Advanced Materials Lecture 5'?"* combines affiliation, authorship, and citation relationships across entities. Here, answering requires traversing a sequence of links. Two possible hypothetical paths are:

$$\text{MIT} \xleftarrow{\text{affiliation}} \text{Researcher} \xrightarrow{\text{published}} \text{Nano2023} \xleftarrow{\text{cites}} \text{Advanced Materials Lecture 5}$$

alternative,

$$\text{Advanced Materials Lecture 5} \xrightarrow{\text{cites}} \text{Nano2023} \xrightarrow{\text{published by}} \text{MIT} \xrightarrow{\text{employs}} \text{Researcher}$$

HyP-KGRAG dynamically adjusts the path length based on question complexity. To support the generation of such paths, a structured prompt is defined to guide the LLM. This prompt instructs the

model to generate step-by-step logical triples for knowledge-grounded retrieval. The full prompt is available in our GitHub repository[1].

Once a query's hypothetical paths are generated, each triple is used to perform semantic similarity search against the offline-constructed vector database to retrieve relevant factual triples. After deduplication, the retrieved factual triples represent the preliminary result set and are passed on to the denoising phase (see Section 3.3).

3.3. LLM-as-Judge Denoising and Generation

Adding all retrieved triples directly into the LLM's input can be problematic due to input size limitations and the potential inclusion of irrelevant or noisy triples retrieved via vector similarity. To address this, HyP-KGRAG incorporates an LLM-as-Judge denoising mechanism inspired by the multi-dimensional evaluation strategies used in GraphRAG and LightRAG. In this process, the LLM scores each candidate triple based on the following three criteria, filtering out low-quality or off-topic information before answer generation.

- *Directness* - refers to how clearly a triple indicates the answer to a query. Triples with relevant entities and relations that directly point to the answer should receive higher scores.
  Scoring: 2 = direct mention of relevant entities/relations; 1 = indirect reference.
- *Entity clarity* - reflects the precision and unambiguity of the entities within a triple, which strongly influences its overall quality.
  Scoring: 2 = explicit entities; 1 = ambiguous references.
- *Context relevance* - captures the LLM's ability to assess whether the answer thoroughly addresses all aspects and details of the question.
  Scoring: 2 = core relevance; 1 = peripheral relevance.

Only triples with a total score of 3 or higher are retained. Triples with ambiguous entities or lacking contextual grounding are excluded, ensuring that only the most relevant and precise information supports the generation process. The full prompt guiding this filtering process is available in our GitHub repository: https://github.com/Zhaotai924/HyP-KGRAG.

After denoising, the filtered and retrieved triples, serving as core contextual information, are combined with the original query and input into the LLM (DeepSeek-V3) to generate a response.

4. Experiments

In this section, the experiments conducted to evaluate the performance of the proposed approach are presented. The source code and the datasets are made publicly available at https://github.com/Zhaotai924/HyP-KGRAG.

4.1. Datasets

The MSE-KG[2] (Material Science and Engineering KG), which represents heterogeneous research data within the NFDI-MatWerk[3] community and beyond, is the dataset used for the experiments.The KG covers three main domains: i) research community structure, including researchers, projects, universities, and institutions; ii) scientific infrastructure, such as software, workflows, instruments, facilities, and educational materials; and iii) research data, spanning repositories, publications, datasets, and reference data. The graph consists of 5,153 triples, 1,533 entities, and 109 distinct relations. To evaluate the proposed approach, 37 competency questions were used. 13 of these questions are complex multi-hop questions.

[1] https://github.com/Zhaotai924/HyP-KGRAG
[2] https://demo.fiz-karlsruhe.de/matwerk/
[3] https://nfdi-matwerk.de/

4.2. Baseline Models

HyP-KGRAG is compared against four baselines:

- **Deepseek V3**: This model operates without utilizing any information from the KG, relying exclusively on the knowledge contained within the Deepseek V3.
- **KGMistral**: This model utilizes SPARQL to extract information from a KG to improve the QA performance of the Mistral model.
- **BM25 Retrieval + Pre-Retrieval Query Processing:** This model first verbalizes the triples and performs sentence-level and paragraph-level sparse retrieval-based matching using BM25. It can also use LLM to process the original query to facilitate keyword-based matching in BM25.
- **BGE-M3 Retrieval + Pre-Retrieval Query Processing**: This model first verbalizes the triples and performs sentence-level and paragraph-level dense retrieval-based matching using BGE-M3. It can also use LLM to process the original query to facilitate semantic matching in BGE-M3.

4.3. Evaluation Metrics and Experiment Setting

The evaluation uses text-matching metrics at three distinct levels: ROUGE-1, ROUGE-2, and ROUGE-L, to provide a systematic assessment for lexical coverage (uni- gram), phrase structure (bigram), and semantic coherence (LCS). Every ROUGE metric assesses several attributes, including ground truth coverage, generation accuracy, and global balance by means of Recall, Precision, and F1 Score, respectively. In addition, BLEU and SBERT similarity measures are used. BLEU provides stricter n-gram precision evaluation, which is especially valuable for measuring phrase-level fidelity. The SBERT similarity measure is based on the cosine similarity computed between sentence embeddings using the lightweight model "all-MiniLM-L6-v2"[19, 20], which offers a strong balance of semantic representation ability and computational efficiency.

For the hyperparameter settings of retrieving document chunks or triples, without loss of generality, all retrievals are configured to return the top-k most relevant document chunks or triples for each query (whether rewritten or not), where k ranges from 10 to 100, sorted by similarity score. These retrieved documents are then fed into the LLM to generate answers using a standardized prompt structure. The DeepSeek-V3 model is the LLM chosen for the experiments conducted in this work.

4.4. Results

As shown in Table 1, HyP-KGRAG consistently outperforms the baseline models across all precision-focused metrics. It achieves the highest F1 scores for ROUGE-1 (0.532), ROUGE-2 (0.408), and ROUGE-L (0.491), along with leading precision score of 0.569, 0.453, and 0.528 respectively. HyP-KGRAG also scores the top BLEU score (0.231), reflecting its strong capability in producing lexically accurate and concise answers. These results are further visualized in Figure 2, where HyP-KGRAG demonstrates superior performance in generating highly accurate and lexically precise answers, significantly outperforming other models in minimizing errors and aligning with reference outputs. While BGE-M3 demonstrates competitive recall, HyP-KGRAG's precision-centric design ensures superior reliability in scenarios demanding factual correctness. For instance, its ROUGE-1 precision (0.569) surpasses BGE-M3's (0.473) by 20%, reflecting stricter adherence to relevant content. Similarly, its BLEU score (0.231) outperforms BGE-M3 (0.200), further validating its lexical alignment strength.

Compared to the earlier KGMistral7B baseline, which integrates SPARQL-based KG retrieval with Mistral-7B generation, HyP-KGRAG achieves substantial improvements, particularly in precision and semantic coherence. Meanwhile, the standalone DeepSeek-V3 model struggles severely on domain-specific datasets (e.g., BLEU=0.002, SBERT=0.283), underscoring its limitations in the absence of prior knowledge or retrieval augmentation.

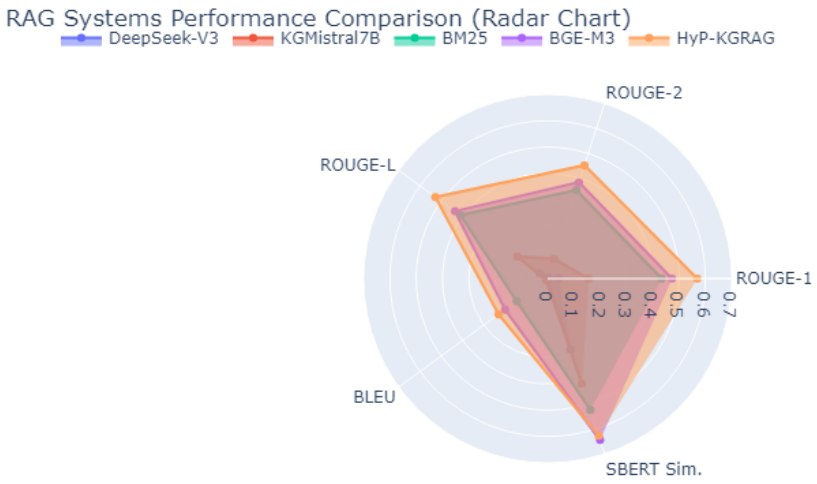

Figure 2: The comparison of the proposed HyP-KGRAG approach against the baseline models using Radar Chart.

4.5. Ablation Studies

This section presents the ablation studies in detail. First, the proposed approach is evaluated by comparing the linear-verbalization technique with LLM-based verbalization. Then, the impact of the denoising module on the overall performance of HyP-KGRAG is assessed.

4.5.1. Verbalization in Hypothetical Path Generation

HyP-KGRAG encodes both the KG and the hypothetical paths using structured triples, represented through a linear verbalization technique as described in Section 3.1. To evaluate the impact of this design choice, an ablation study is conduct in which the linear-verbalization is replaced with sentence-level natural language expressions generated by an LLM. Note that the sentence-level verbalization is chosen here because it demonstrated a better performance as compared to paragraph-level verbalization for the baseline models.

As shown in Table 2, the results demonstrate that triple-based hypothetical path matching, i.e., applying linear-verbalization, achieves significantly higher retrieval accuracy than its natural language based verbalization. This difference highlights the advantages of structured representations in RAG systems: triples offer compact, unambiguous, and noise-resistant semantic cues that are better aligned with the query intent.

In contrast, natural language representations, while richer in context, introduce higher variability and redundancy. This leads to less effective matching during retrieval, particularly in precision-sensitive tasks. Overall, the results support the effectiveness of the linear verbalization approach in closing the semantic gap between queries and structured knowledge, thus reinforcing its role as a core design choice in HyP-KGRAG.

4.5.2. Effect of Denoising Removal

This ablation study evaluates the impact of removing the LLM-based denoising mechanism from HyP-KGRAG. As shown in Table 3, denoising plays a critical role in enhancing both retrieval precision and the overall quality of generated responses. Across all precision-related metrics, denoising consistently leads to notable improvements indicating that noisy or irrelevant triples are effectively filtered out, allowing for more targeted and accurate content generation. One of the most noticeable improvements

Table 1
Performance comparison of HyP-KGRAG against baseline models across lexical (ROUGE, BLEU) and semantic (SBERT) evaluation metrics. For the BM25 and BGE-M3 baselines, both sentence-level and paragraph-level verbalization strategies were explored; however, only sentence-level results are reported here, as they consistently outperformed paragraph-level. Sentence-level verbalization converts each triple into an independent natural language sentence, treated as a separate document chunk. Paragraph-level verbalization, given an entity, aggregates all triples where the entity appears as the subject and generates a natural language paragraph using an LLM. Paragraph-level results are not included in this table.

RAG Systems	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	SBERT Sim.
DeepSeek-V3					
F1	0.037	0.004	0.033		
Precision	0.042	0.006	0.038	0.002	0.283
Recall	0.068	0.005	0.060		
KGMistral7B[16]					
F1	0.150	0.071	0.134		
Precision	0.156	0.079	0.143	0.008	0.420[21]
Recall	0.245	0.120	0.225		
BM25					
F1	0.406	0.338	0.383		
Precision	0.434	0.354	0.410	0.146	0.525
Recall	0.518	0.422	0.486		
BGE-M3					
F1	0.462	0.361	0.425		
Precision	0.473	0.385	0.436	0.200	**0.645**
Recall	**0.584**	**0.460**	**0.542**		
HyP-KGRAG					
F1	**0.532**	**0.408**	**0.491**		
Precision	**0.569**	**0.453**	**0.528**	**0.231**	0.629
Recall	0.575	0.453	0.529		

Table 2
Performance comparison between HyP-KGRAG and its verbalization variant. The last column shows the relative decrease in performance.

Metric	HyP-KGRAG	HyP-KGRAG+Sentence-Level-Verbalization	Decrease (%)
ROUGE-1-F	0.532	0.409	-23.1%
ROUGE-2-F	0.414	0.310	-25.1%
ROUGE-L-F	0.491	0.381	-22.4%
BLEU	0.231	0.170	-26.4%
SBERT	0.629	0.585	-7.0%

is the BLEU score, which increases by 16.5%, confirming a stronger lexical alignment between the generated output and the ground truth. This suggests that denoising enhances the model's ability to focus on essential content and reduce spurious or off-topic information.

However, this gain in precision comes with a slight trade-off. Specifically, ROUGE-1 recall drops by 4.2%, pointing to a potential loss of some relevant details due to the aggressive filtering of context. Despite this, the F1 scores across ROUGE and BLEU remain consistently positive, demonstrating an effective balance between precision and recall. Importantly, the semantic coherence of the outputs remains intact. The SBERT similarity score shows negligible change (from 0.598 to 0.596), indicating that the high-level semantic alignment between the generated response and the reference answer is

preserved even when some lexical content is pruned. In summary, this analysis confirms that the denoising mechanism in HyP-KGRAG serves as a powerful tool for improving precision and refining generation quality, without compromising the underlying conceptual integrity of the output.

Table 3
Evaluation results before and after denoising.

Metric	Before Denoising (Avg)	After Denoising (Avg)	Trend
ROUGE-1-F1	0.459	0.486	↑ +5.8%
ROUGE-1-Precision	0.491	0.534	↑ +8.7%
ROUGE-1-Recall	0.543	0.520	↓ -4.2%
ROUGE-2-F1	0.354	0.382	↑ +7.8%
ROUGE-2-Precision	0.383	0.421	↑ +9.9%
ROUGE-2-Recall	0.421	0.405	↓ -3.8%
ROUGE-L-F1	0.416	0.454	↑ +9.3%
ROUGE-L-Precision	0.446	0.500	↑ +12.1%
ROUGE-L-Recall	0.492	0.488	↓ -0.7%
BLEU	0.172	0.200	↑ +16.5%
SBERT Similarity	0.598	0.596	≈ Stable

5. Conclusion and Future Work

5.1. Summary

This work introduces HyP-KGRAG, a knowledge-grounded RAG framework that redefines how structured information from KGs is utilized in domain-specific question answering. By directly operating on triples, rather than relying on potentially noisy natural language verbalizations, HyP-KGRAG avoids linguistic ambiguities and better captures the intent. The incorporation of hypothetical path generation allows for dynamic multi-hop evidence retrieval, while the hybrid retrieval strategy (BM25 + BGE-M3) ensures high-recall access to relevant knowledge. Additionally, the LLM-as-judge denoising module significantly boosts answer quality by filtering irrelevant or noisy triples prior to generation.

Experimental results on the MSE-KG dataset validate the effectiveness of our design, showing strong gains in both factual precision (e.g., ROUGE-1 F1: 0.532) and semantic coherence (SBERT similarity: 0.629) over established baselines. These findings highlight the promise of structured, triple-based RAG approaches in enhancing reliability and accuracy in knowledge-intensive tasks.

5.2. Limitations

Key limitations include: i) the system's effectiveness still depends significantly on the reasoning capabilities of the underlying LLM and ii) current evaluation is restricted to a single dataset, i.e. MSE-KG, which may limit generalizability.

5.3. Future Work

Future work will focus on improving the scalability and generalizability of HyP-KGRAG by extending evaluations to diverse multi-hop KGQA benchmarks. This will allow for a more comprehensive assessment of the model's robustness and applicability. Additionally, the integration of dynamic path retrieval mechanisms that more closely reflect ground-truth graph structures could further enhance multi-hop evidence retrieval quality. Another promising direction is the incorporation of multimodal knowledge through visual-language KGs and image-attributed triples, enabling image-augmented RAG (iRAG) models that improve contextual grounding and reduce hallucinations. The development of agentic capabilities where the system can autonomously select and sequence retrieval, rewriting, and verification modules may also lead to more adaptive and reliable performance. While the current

evaluation primarily focuses on addressing hallucination challenges in domain-specific QA with the DeepSeek LLM, future work will involve more extensive experiments comparing HyP-KGRAG with state-of-the-art models such as GraphRAG and LightRAG to more rigorously assess its comparative performance.

References

[1] K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, M. Amin, L. Hou, K. Clark, S. R. Pfohl, H. Cole-Lewis, et al., Toward expert-level medical question answering with large language models, Nature Medicine (2025) 1–8.

[2] T. Zhang, F. Ladhak, E. Durmus, P. Liang, K. McKeown, T. B. Hashimoto, Benchmarking large language models for news summarization, Transactions of the Association for Computational Linguistics 12 (2024) 39–57.

[3] L. Wang, C. Lyu, T. Ji, Z. Zhang, D. Yu, S. Shi, Z. Tu, Document-level machine translation with large language models, arXiv preprint arXiv:2304.02210 (2023).

[4] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, Advances in neural information processing systems 33 (2020) 9459–9474.

[5] A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, et al., Deepseek-v3 technical report, arXiv preprint arXiv:2412.19437 (2024).

[6] S. Robertson, H. Zaragoza, et al., The probabilistic relevance framework: Bm25 and beyond, Foundations and Trends® in Information Retrieval 3 (2009) 333–389.

[7] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, Z. Liu, Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024. arXiv:2402.03216.

[8] J. Baek, A. F. Aji, A. Saffari, Knowledge-augmented language model prompting for zero-shot knowledge graph question answering, arXiv preprint arXiv:2306.04136 (2023).

[9] R.-C. Chang, J. Zhang, Communitykg-rag: Leveraging community structures in knowledge graphs for advanced retrieval-augmented generation in fact-checking, arXiv preprint arXiv:2408.08535 (2024).

[10] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, D. Metropolitansky, R. O. Ness, J. Larson, From local to global: A graph rag approach to query-focused summarization, arXiv preprint arXiv:2404.16130 (2024).

[11] Z. Guo, L. Xia, Y. Yu, T. Ao, C. Huang, Lightrag: Simple and fast retrieval-augmented generation (2024).

[12] L. Luo, Y.-F. Li, G. Haffari, S. Pan, Reasoning on graphs: Faithful and interpretable large language model reasoning, arXiv preprint arXiv:2310.01061 (2023).

[13] C. Mavromatis, G. Karypis, Gnn-rag: Graph neural retrieval for large language model reasoning, arXiv preprint arXiv:2405.20139 (2024).

[14] X. Zhu, Y. Xie, Y. Liu, Y. Li, W. Hu, Knowledge graph-guided retrieval augmented generation, arXiv preprint arXiv:2502.06864 (2025).

[15] L. Liang, Z. Bo, Z. Gui, Z. Zhu, L. Zhong, P. Zhao, M. Sun, Z. Zhang, J. Zhou, W. Chen, et al., Kag: Boosting llms in professional domains via knowledge augmented generation, in: Companion Proceedings of the ACM on Web Conference 2025, 2025, pp. 334–343.

[16] M. Li, H. Yang, Z. Liu, M. M. Alam, H. Sack, G. A. Gesese, et al., Kgmistral: Towards boosting the performance of large language models for question answering with knowledge graph integration, in: Workshop on Deep Learning and Large Language Models for Knowledge Graphs, 2024.

[17] L. Gao, X. Ma, J. Lin, J. Callan, Precise zero-shot dense retrieval without relevance labels, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 1762–1777.

[18] J. Johnson, M. Douze, H. Jégou, Billion-scale similarity search with gpus, IEEE Transactions on Big Data 7 (2019) 535–547.
[19] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, M. Zhou, Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, arXiv preprint arXiv:2002.10957 (2020).
[20] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, arXiv preprint arXiv:1908.10084 (2019).
[21] Mingze, Kgmistral, https://github.com/Mingze101/KGMistral/, 2023. Accessed: 2025-05-26.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT in order to: Grammar and spelling check. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.