Mindful-RAG: A Study of Points of Failure in Retrieval Augmented Generation

Garima Agrawal, Tharindu Kumarage, Zeyad Alghamdi, and Huan Liu
School of Computing and Augmented Intelligence, Arizona State University, Tempe, USA
Email: {garima.agrawal, kskumara, zalgham1, huanliu}@asu.edu

***Abstract*—Large Language Models (LLMs) excel at generating coherent text but often struggle with knowledge-intensive queries, particularly in domain-specific and factual question-answering tasks. Retrieval-augmented generation (RAG) systems have emerged as a promising solution by integrating external knowledge sources, such as structured knowledge graphs (KGs). While KG-based RAG approaches have demonstrated value, current state-of-the-art solutions frequently fall short, failing to deliver accurate and reliable answers even when the necessary factual knowledge is available. In this paper, we present a critical analysis of failure points in existing KG-based RAG methods, identifying eight key areas of concern, including misinterpretation of question context, incorrect relation mapping, and ineffective ambiguity resolution. We argue that these failures primarily stem from design limitations in current KG-RAG systems, such as inadequate attention to discerning user intent and insufficient alignment of retrieved knowledge with the contextual demands of the query. Based on this analysis, we propose a new approach for KG-RAG systems, termed Mindful-RAG, which re-engineers the retrieval process to be more intent-driven and contextually aware. By enhancing reasoning capabilities, improving constraint identification, and addressing the structural limitations of knowledge graphs, we aim to improve the reliability and effectiveness of KG-RAG systems. To validate this approach, we developed a proof-of-concept by integrating the principles of Mindful-RAG into an existing KG-RAG system. The Mindful-RAG approach seeks to deliver more robust, accurate, and contextually aligned AI-driven knowledge retrieval systems, with potential applications in critical domains such as healthcare, legal, research, and scientific discovery, where precision and reliability are paramount.**

***Index Terms*—LLMs, Knowledge Graphs (KG), Retrieval Augmented Generation (RAG), Hallucinations, Points of Failure**

I. Introduction

Large Language Models (LLMs) have revolutionized natural language processing, excelling in various tasks. However, they frequently generate hallucinated responses when dealing with domain-specific or knowledge-intensive queries [1]. This limitation has led to the development of Retrieval-augmented Generation (RAG) methods, which enable LLMs to access and incorporate external knowledge sources, such as structured knowledge graphs (KGs) [2], [3].

Despite the promise of RAG methods, particularly those integrating KGs, significant challenges persist. Even with access to relevant information, these systems often fail to provide accurate answers, especially as query complexity increases [4]–[6]. To better understand these limitations, We conducted a study to critically analyze the failure points in existing KG-based RAG methods. Our investigation identified eight critical failure points in these systems, which we categorized into two primary areas:

1) **Reasoning Failures**: LLMs struggle to accurately interpret user queries and leverage contextual information, resulting in a misalignment between retrieved knowledge and query intent.
2) **Structural Limitations**: These failures primarily arise from insufficient attention to the structure of knowledge sources, such as knowledge graphs, and the use of inappropriate evaluation metrics.

Given these persistent issues, there is a pressing need to look beyond conventional approaches and critically reassess how KG-RAG systems are designed. To address these challenges, we propose **Mindful-RAG**, an approach that re-engineers the retrieval process to be more intent-driven and contextually aware. Mindful-RAG is not merely an alternative method; it represents a comprehensive approach aimed at the development of more effective KG-RAG systems. Unlike traditional methods that primarily rely on semantic similarity or structural cues, Mindful-RAG suggests to leverage the intrinsic parametric knowledge of LLMs to accurately discern the intent behind queries. This approach not only guides the retrieval process to ensure that the extracted context from the KG is relevant but also aligns it with the original intent of the query. Additionally, Mindful-RAG introduces advanced contextual alignment techniques for efficient knowledge graph navigation and incorporates a validation step to ensure the generated response meets the intended requirements. To validate this approach, we developed a proof-of-concept that integrates Mindful-RAG into an existing KG-RAG system through prompt engineering. Preliminary experiments on WebQSP and MetaQA datasets indicate promising results compared to existing state-of-the-art methods, particularly in reducing reasoning errors by enhancing the focus on understanding query objectives and improving contextual alignment. In this paper, we make the following key contributions:

1) A comprehensive error analysis of KG-based RAG methods, identifying eight critical failure points and highlighting design limitations in state-of-the-art frameworks, particularly in addressing question intent and achieving contextual alignment in vanilla-RAG systems.
2) The introduction of a proof-of-concept that opens a novel research direction, redefining the RAG pipeline by leveraging LLMs' parametric memory for enhanced

intent identification and contextual alignment.

By addressing these fundamental challenges, Mindful-RAG aims to create more robust, accurate, and contextually aligned AI-driven knowledge retrieval systems, particularly in precision-critical fields such as healthcare, legal, research, and scientific discovery.

II. KG-based RAG Failure Analysis

Various methodologies have been developed to enhance LLMs with KG-based RAG systems. By leveraging structured and meticulously curated knowledge from these graphs, the retrieved information is more likely to be factually accurate.

We assessed the effectiveness of these methods and analyzed their accuracy in retrieving information for fact-based question-answering (QA) tasks using a KG. Although most of these models surpass the performance of zero-shot QA conducted directly from various standard LLMs, there is still considerable scope for improvement. For our study, we chose the WebQuestionsSP (WebQSP) [7] dataset for knowledge graph question answering (KGQA), which is frequently utilized by KG-based RAG methods [8]. This dataset, based on the Freebase KG [9], consists of questions that require up to two or three-hop reasoning to identify the correct answer entity, utilizing Hits@k as the evaluation metric to determine if the top-k predicted answer is accurate. It includes approximately 1600 test samples. The vanilla ChatGPT (GPT-3.5) accuracy in zero-shot setting without any external knowledge is 61.2%.

StructGPT [10] is a state-of-the-art approach that leverages LLM's capabilities for reasoning with evidence extracted from a KG. This method involves extracting a sub-graph from a KG by matching the topic entities in the question. The LLM is then directly employed to identify useful relations and extract relevant triples from the sub-graph, guiding it to effectively traverse and reason within the graph structure. The Hits@1 accuracy of StructGPT on the WebQSP dataset, when utilizing ChatGPT (GPT-3.5) for question-answering tasks, was reported to be 72.6%. In this study, we have chosen StructGPT as our reference model to analyze the current SOTA developments of KG-based RAGs in the QA setting.

We began our analysis by closely examining the failure instances of StructGPT on the WebQSP dataset. We meticulously reviewed logs from around 435 error cases to understand the model's behavior during the reasoning process. Initially, we manually analyzed 10% of these error samples to identify common error types. Building on this manual analysis, we developed an LLM-assisted pipeline using few-shot samples to categorize the remaining errors and identify any additional error types. We then employed LLM-critic to provide recommendations and identify recurring themes based on our analysis and category mapping. These LLM-generated suggestions were subsequently validated through manual review. This detailed examination allowed us to pinpoint distinct error patterns, leading to the identification of eight primary error categories. These issues were further organized into two main divisions: *Reasoning Failures*, which encompass errors arising from reasoning deficiencies, and *Structural Limitations*, which include structural issues within the knowledge graph.

Reasoning Failures: Most failures stem from the LLMs' inability to reason correctly. These issues primarily include a failure to accurately understand the question, leading to difficulty in mapping the question to the available information. Additionally, LLMs struggle to effectively apply the cues in the question to narrow down the relevant entities. They also often fail to apply specific constraints that logically limit the search space. Generally, LLMs have difficulty grasping specifics such as temporal context, aggregating or summarizing answers, and disambiguating among multiple choices. Furthermore, they frequently choose incorrect relations, particularly in complex queries requiring multi-hop reasoning, finding it challenging to focus on the relevant elements necessary to formulate an answer. In Table I, we detail various reasoning failures, each illustrated with an example.

Structural Limitations: These issues occur when knowledge becomes inaccessible due to limitations in the structural design of the knowledge base or inefficient processing methods. In Table I, we categorize these challenges under structural issues, limited query processing, and the selection of inappropriate evaluation metrics.

In this work, our primary focus is on addressing errors stemming from reasoning failures in LLM models and enhancing their reasoning capabilities. Structural limitations, on the other hand, can be resolved through careful programming and the selection of more appropriate evaluation metrics. Our analysis of reasoning error samples reveals two main challenges: **(i)** Models frequently fail to grasp the question's intent, relying primarily on structural cues and semantic similarity to extract relevant relations and generate answers. **(ii)** They struggle to align the question's context with the available information.

This inability to comprehend intent and context leads to incorrect relation rankings and the misapplication of constraints. A review of response logs from both failed and successful interactions reveals that the LLM relies heavily on semantic matching. While this approach suffices for simple queries, it falls short in handling complex questions that demand multi-hop reasoning and deep contextual understanding. Therefore, improving intent identification and context alignment is essential for enhancing model performance.

III. Mindful-RAG

In response to our findings, we introduce **Mindful-RAG**, designed to address two critical gaps: the lack of question-intent identification and the insufficient contextual alignment with available knowledge. This approach employs a strategic hybrid method that integrates the model's intrinsic parametric knowledge with non-parametric external knowledge from a KG. The following steps provide a detailed overview of our design and methodology, each accompanied by an illustrative example.

- **Step 1. Identify key Entities and relevant Tokens:** The first step is to pinpoint the key entities within a question to facilitate the extraction of pertinent information from

Table I
KG-BASED RAG FAILURE ANALYSIS

Error Category	Error Type	Description	Representative Failed Example(s)
Reasoning Failures	Misinterpretation of Question's Context	LLMs misinterpret the question or fail to understand specific requirements of the question.	• Failed to relate **Justin Bieber's** birthplace to his country of birth, focusing on city-level information instead of the required higher geographical context. • Incorrectly identified the location of **Fukushima Daiichi** nuclear plant, choosing the city **'Fukushima'** instead of the correct town **'Okuma'** and country **'Japan'**.
	Incorrect Relation Mapping	LLMs often choose relations that do not correctly address the question.	For a question about where **Andy Murray** started playing tennis, choosing *people.person.place_of_birth* suggests a misunderstanding of the question's intent.
	Ambiguity in Question or Data	LLMs fail to identify key terms and their meanings or implications across various contexts from the provided KG triples.	• Could not identify the **Serbian language** from the list of languages spoken in **Serbia**. • Failed to recognize that a query was about the **"most" exported item**, not just any exported item.
	Specificity or Precision Errors	LLMs often misinterpret questions requiring aggregated responses as specific, singular answers. They also struggle with temporal context.	• Picked **2000** as **George W. Bush's** election year without considering his two elections (**2000 and 2004**). • Selected **'Sue Douglas'** as **Niall Ferguson's** spouse instead of finding the current spouse, **'Ayaan Hirsi Ali'**, ignoring multiple spouse possibilities.
	Constraint Identification Error	LLMs fail to correctly identify or apply constraints provided or implied in the question.	• Could not effectively narrow the search for **Jackie Robinson's** first team. • For "Who played Bilbo in Lord of the Rings?", LLMs identified **"Old Bilbo"** and specific films but failed to derive a single definitive answer.
Structural Limitation	Encoding Issues	Compound value types (CVTs) in KGs represent complex data. If mismanaged or unrecognized by LLMs, they may be misinterpreted as final answers.	For 'Where is the Sony Ericsson Company?', the model correctly identifies relations but mistakenly selects CVT_0 as the final answer due to CVT node linking.
	Inappropriate Evaluation	The exact match (EM) module only accepts fully correct answers and sometimes fails due to misinterpreting the required depth of information or mis-aligning with expected answer format.	For 'What year did the Orioles go to the World Series?', the model retrieved correct years (1983, 1970, 1966) but failed to match the expected format **[1983 World Series, 1970 World Series, 1966 World Series]**.
	Limited Query Processing	Instances where the model recognizes that further information is required for a conclusive answer, yet receives no feedback, indicating a gap in programming or query processing.	The model responds **'Need More Information'** for 'What is the name of the San Francisco newspaper?' but receives no further feedback.

an external KG or a sub-graph within a KG. Additionally, in our method, we task the LLM model with identifying other significant tokens that may be crucial for answering the question. For instance, consider the question from WebQSP, **"Who is Niall Ferguson's wife?"** The key entity identified by the model is *'Niall Ferguson'*, and the other relevant token is *'wife'*.

- **Step 2. Identify the Intent:** In this step, we leverage the LLM's understanding to discern the intent behind the question, prompting it to focus on keywords and phrases that clarify the depth and scope of the intent. For instance, in the provided example, the model identifies the question's intent as *"identify spouse"*.
- **Step 3. Identify the Context:** Next, the model was instructed to understand and analyze the context of the question, which is essential for formulating an accurate response. For the provided example, the model identifies relevant contextual aspects such as *"personal relationships," "marital status," and "current spouse."*
- **Step 4. Candidate Relation Extraction:** Next, the key entity relations are extracted from the sub-graph within one-hop distance. For our example, the candidate relations include information about the subject's profession, personal life, and societal role.
- **Step 5. Intent-Based Filtering and Contextual Ranking of Relations:** In this step, the model conducts a detailed analysis to filter and rank the extracted relations and entities based on the question's intent, ensuring

relevance and accuracy. Relations are ranked according to their contextual significance, with the top-k relations being selected. For example, considering the intent and context in the given scenario, the model identifies *"people.person.spouse_s"* as the most relevant relation.

- **Step 6. Contextual Alignment of Constraints:** In this step, the model considers temporal and geographical constraints by utilizing relevant data from various indicators to address more complex queries. This process ensures that responses are accurately tailored to specific times, locations, or historical periods. Once constraints are identified, the model aligns them contextually and refines the list of candidate entities. For example, in our scenario, the model identified constraints such as names of spouses, marriage start and end times, and the location of the ceremony. It then narrowed the list to potential spouses and extracted all related triples. Finally, the model aligned this information with the context of the 'current spouse,' resulting in the correct response of *'Ayaan Hirsi Ali'*, in contrast to existing methods [10], where the LLM incorrectly selected the first name on the spouse list, *'Sue Douglas'*.
- **Step 7. Intent-Based Feedback:** In the final step, the model is prompted to validate whether the final answer aligns with the initially identified intent and context of the question. If the answer does not meet these criteria, the model is instructed to revisit **Step 5 and 6** to further refine its response.

Similarly, the model adeptly contextualizes and aggregates pertinent information in other instances. For example, when asked, *"What songs did Justin Bieber write?"* it successfully compiles all relevant songs. In response to, *"What is the state flower of Arizona?"* it identifies ***'Arizona'*** as the key entity, with ***'state'*** and ***'flower'*** as relevant tokens. It correctly interprets the intent to ***"identify state flower"*** and recognizes the context of ***'botany,' 'state symbols,'*** and ***'Arizona's official flora'*** choosing the appropriate relation: *"government.governmental_jurisdiction.official_symbols."* In contrast, traditional methods only identify *'Arizona'* as the key entity, often missing the broader context, leading to choosing incorrect relations, *"base.locations.states_and_provinces.country"* and answer stating the state flower of Arizona is unknown.

Mindful-RAG leverages the LLM's intrinsic understanding in the first three steps to identify not only the key entities but also to gather additional information such as relevant tokens, intent, and current context, all of which are essential for accurately answering the question. These steps enable the model to appropriately filter relations and align constraints with the current context. By incorporating these steps, the LLM becomes more mindful of the specific elements to consider. In the final two steps, the LLM is prompted to tailor its response and align it with specific constraints such as time, location, and any requirements for aggregating an answer.

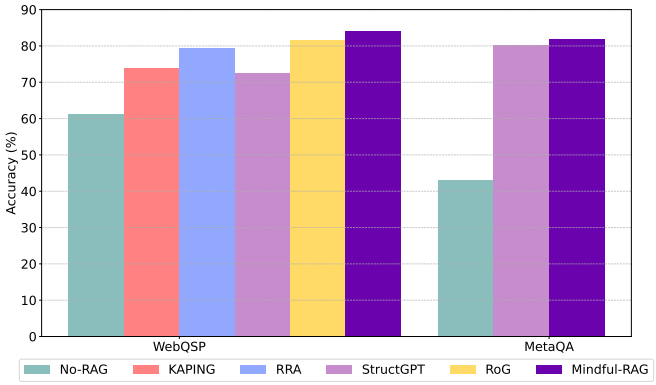

Figure 1. Mindful-RAG results on WebQSP and MetaQA

IV. Experiments and Results

Datasets: We evaluate our approach on two benchmark KGQA datasets, specifically WebQSP and MetaQA [11]. MetaQA features questions related to the movie domain, with answers up to three hops away from the topic entities in a movie KG (based on OMDb). Here, we focused only on 3-hop questions.

In our analysis of the WebQSP dataset, we evaluated several baseline methods: KAPING [12], Retrieve-Rewrite-Answer (RRA) [13], Reasoning on Graphs (RoG) [14], and StructGPT [10]. For MetaQA (3-hop), StructGPT [10] served as the baseline. The results for these methods were taken directly from the respective publications. In our experiments, we adapted the base code of StructGPT [10] and modified it only for improved reasoning as outlined in the previous section. We also examined the performance of ChatGPT without RAG on both datasets. The results, presented in Figure 1, show that Mindful-RAG, shows promising improvement in accuracy of reasoning error cases achieving a Hits@1 of 84% on WebQSP and 82% on MetaQA. Additional accuracy improvements can be achieved by addressing structural issues and incorporating partial answers to enhance precision, rather than relying solely on exact matches.

The primary goal of this study is to explore methods for mitigating reasoning errors in KG-RAG systems. It is important to emphasize that our approach serves as an initial demonstration of the potential in combining the parametric knowledge of models with non-parametric external knowledge. More advanced RAG methods could be developed in the future to significantly surpass the performance of our approach.

V. Related Work

Recent efforts to enhance RAG systems have focused on various improvements. Siriwardhana et al. [15] aimed to improve domain adaptation for Open Domain Question Answering (ODQA) by jointly training the retriever and generator and enriching the Wikipedia-based knowledge base with healthcare and news content. RAFT [16] enhances RAG by customizing language models for specific domains in open-book QA. Self-RAG [17] aims to increase the factual accuracy of LLMs

through adaptive self-critique and retrieval-generation feedback loops. Fit-RAG [18] introduces a method that uses detailed prompts to ensure deep question understanding and clear reasoning in fact retrieval. Domain-specific knowledge graphs [19]–[22] have been effectively employed in KG-based RAG systems within LLMs [23]–[25] for question-answering tasks [26], [27]. While most efforts focus on enhancing LLMs by augmenting knowledge graphs with relevant facts, there has been limited work on improving the reasoning capabilities of LLMs during knowledge retrieval. Our research with Mindful-RAG aims to establish a road map for advancing these methods by leveraging the model's inherent knowledge for better question understanding.

VI. Discussion and Conclusion

We conducted an error analysis of KG-based RAG methods integrated with LLMs for question-answering tasks, identifying eight critical failure points, categorized into reasoning failures and structural limitations. Reasoning failures involve LLMs struggling with understanding questions and leveraging contextual clues, particularly in cases involving temporal context and complex relational reasoning. Structural limitations pertain to inadequate attention to the structure of the knowledge base and weaknesses in evaluation metrics. These challenges highlight areas for improvement, especially in handling complex, multi-hop queries. To address these issues, we propose Mindful-RAG, designed to enhance intent-driven retrieval and ensure contextually coherent responses, directly targeting the identified deficiencies. While our approach focuses on mitigating reasoning-based failures, future research could explore addressing structural issues through the use of feedback and human-in-the-loop input. Additionally, combining vector-based search with KG-based sub-graph retrieval represents a promising direction for enhancing LLM performance in knowledge-intensive tasks.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 2335666.

References

[1] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.

[2] J. Li, Y. Yuan, and Z. Zhang, "Enhancing llm factual accuracy with rag to counter hallucinations: A case study on domain-specific queries in private knowledge-bases," *arXiv preprint arXiv:2403.10446*, 2024.

[3] Y. Ding, W. Fan, L. Ning, S. Wang, H. Li, D. Yin, T.-S. Chua, and Q. Li, "A survey on rag meets llms: Towards retrieval-augmented large language models," *arXiv preprint arXiv:2405.06211*, 2024.

[4] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, and H. Wang, "Retrieval-augmented generation for large language models: A survey," *arXiv preprint arXiv:2312.10997*, 2023.

[5] G. Agrawal, T. Kumarage, Z. Alghami, and H. Liu, "Can knowledge graphs reduce hallucinations in llms?: A survey," *arXiv preprint arXiv:2311.07914*, 2023.

[6] S. Jeong, J. Baek, S. Cho, S. J. Hwang, and J. C. Park, "Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity," *arXiv preprint arXiv:2403.14403*, 2024.

[7] W.-t. Yih, M. Richardson, C. Meek, M.-W. Chang, and J. Suh, "The value of semantic parse labeling for knowledge base question answering," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2016, pp. 201–206.

[8] Y. Tan, D. Min, Y. Li, W. Li, N. Hu, Y. Chen, and G. Qi, "Can chatgpt replace traditional kbqa models? an in-depth analysis of the question answering performance of the gpt llm family," in *International Semantic Web Conference*. Springer, 2023, pp. 348–367.

[9] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: a collaboratively created graph database for structuring human knowledge," in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, 2008, pp. 1247–1250.

[10] J. Jiang, K. Zhou, Z. Dong, K. Ye, W. X. Zhao, and J.-R. Wen, "Structgpt: A general framework for large language model to reason over structured data," *arXiv preprint arXiv:2305.09645*, 2023.

[11] Y. Zhang, H. Dai, Z. Kozareva, A. Smola, and L. Song, "Variational reasoning for question answering with knowledge graph," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

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

[13] Y. Wu, N. Hu, G. Qi, S. Bi, J. Ren, A. Xie, and W. Song, "Retrieve-rewrite-answer: A kg-to-text enhanced llms framework for knowledge graph question answering," *arXiv preprint arXiv:2309.11206*, 2023.

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

[15] S. Siriwardhana, R. Weerasekera, E. Wen, T. Kaluarachchi, R. Rana, and S. Nanayakkara, "Improving the domain adaptation of retrieval augmented generation (rag) models for open domain question answering," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1–17, 2023.

[16] T. Zhang, S. G. Patil, N. Jain, S. Shen, M. Zaharia, I. Stoica, and J. E. Gonzalez, "Raft: Adapting language model to domain specific rag," *arXiv preprint arXiv:2403.10131*, 2024.

[17] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "Self-rag: Learning to retrieve, generate, and critique through self-reflection," *arXiv preprint arXiv:2310.11511*, 2023.

[18] Y. Mao, X. Dong, W. Xu, Y. Gao, B. Wei, and Y. Zhang, "Fit-rag: Black-box rag with factual information and token reduction," *arXiv preprint arXiv:2403.14374*, 2024.

[19] B. Abu-Salih, "Domain-specific knowledge graphs: A survey," *Journal of Network and Computer Applications*, vol. 185, p. 103076, 2021.

[20] G. Agrawal, Y. Deng, J. Park, H. Liu, and Y.-C. Chen, "Building knowledge graphs from unstructured texts: Applications and impact analyses in cybersecurity education," *Information*, vol. 13, no. 11, p. 526, 2022.

[21] X. Tang, Z. Feng, Y. Xiao, M. Wang, T. Ye, Y. Zhou, J. Meng, B. Zhang, and D. Zhang, "Construction and application of an ontology-based domain-specific knowledge graph for petroleum exploration and development," *Geoscience Frontiers*, vol. 14, no. 5, p. 101426, 2023.

[22] G. Agrawal, K. Pal, Y. Deng, H. Liu, and C. Baral, "Aiseckg: Knowledge graph dataset for cybersecurity education," *AAAI-MAKE 2023: Challenges Requiring the Combination of Machine Learning 2023*, 2023.

[23] J. Delile, S. Mukherjee, A. Van Pamel, and L. Zhukov, "Graph-based retriever captures the long tail of biomedical knowledge," *arXiv preprint arXiv:2402.12352*, 2024.

[24] X. Jiang, R. Zhang, Y. Xu, R. Qiu, Y. Fang, Z. Wang, J. Tang, H. Ding, X. Chu, J. Zhao *et al.*, "Hykge: A hypothesis knowledge graph enhanced framework for accurate and reliable medical llms responses."

[25] G. Agrawal, K. Pal, Y. Deng, H. Liu, and Y.-C. Chen, "Cyberq: Generating questions and answers for cybersecurity education using knowledge graph-augmented llms," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 21, 2024, pp. 23 164–23 172.

[26] Y. Zhao, Z. Li, and J. Wang, "Lb-kbqa: Large-language-model and bert based knowledge-based question and answering system," *arXiv preprint arXiv:2402.05130*, 2024.

[27] G. Agrawal, D. Bertsekas, and H. Liu, "Auction-based learning for question answering over knowledge graphs," *Information*, vol. 14, no. 6, p. 336, 2023.