SvedbergOpen
DISSEMINATION OF KNOWLEDGE
International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/
Research Paper Open Access

Graph-Augmented Retrieval Framework For Explainable Domain-Specific Large Language Model Reasoning

Yogesh A. Bhavsar[1], Vimal Bibhu[2], Amol D. Wable[3], Bipin Sule[4], Aseem Aneja[5], Vijayshree Khugshal[6], Dr. Shweta Suryawanshi[7], Dr. Lowlesh Nandkishor Yadav[8]

[1] Assistant Professor, Department of Mechanical Engineering, Sanjivani College of Engineering, Savitribai Phule Pune University, Kopargaon 423603, Ahilyanagar, Maharashtra, India.
Email: bhavsaryogeshmech@sanjivani.org.in Email: Yogesh.star1@gmail.com
[2] Department of Computer Science & Engineering, Noida International University, Greater Noida, Uttar Pradesh 203201, India. Email: vimal.bhibu@niu.edu.in
[3] Assistant Professor, Department of Mechanical Engineering, Sanjivani College of Engineering, Savitribai Phule Pune University, Kopargaon 423603, Ahilyanagar, Maharashtra, India. Email: wableamolmech@sanjivani.org.in
[4] Department of Engineering, Science and Humanities, Vishwakarma Institute of Technology, Pune, Maharashtra 411037, India. Email: bipin.sule@vit.edu
[5] Centre for Research Impact & Outcome, Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura 140401, Punjab, India. Email: aseem.aneja.orp@chitkara.edu.in ORCID: 0009-0009-5607-1688
[6] Assistant Professor, Department of Computer Application, Dev Bhoomi Uttarakhand University, Dehradun, Uttarakhand, India. Email: socse.vijayshree@dbuu.ac.in ORCID: 0009-0002-0277-2154
[7] Assistant Professor, Department of Semiconductor Engineering, D. Y. Patil International University, Akurdi, Maharashtra, India. Email: shweta.suryawanshi@dypiu.ac.in
[8] Department of Computer Engineering, Suryodaya College of Engineering and Technology, Nagpur, Maharashtra, India. Email: lowlesh.yadav@gmail.com

ABSTRACT
Large language models have demonstrated strong capabilities in natural language understanding and reasoning; however, their use in domain-specific applications is limited by hallucination, weak factual grounding, and insufficient explainability. This study proposes a graph-augmented retrieval framework that combines semantic vector retrieval with knowledge-graph traversal to support accurate and transparent domain-specific reasoning. The framework retrieves relevant document segments, identifies connected entities and relationships, ranks the combined evidence, and supplies the selected knowledge to the language model for answer and explanation generation. The proposed method is evaluated against a standalone LLM, conventional retrieval-augmented generation, graph-only retrieval, and an existing GraphRAG approach. Performance is assessed using retrieval precision, answer accuracy, explainability score, hallucination rate, and response latency. The results demonstrate that the proposed framework improves evidence relevance, reasoning accuracy, and explanation quality while reducing unsupported responses, with a moderate increase in computational latency. The framework provides a practical foundation for trustworthy LLM deployment in knowledge-intensive domains.

Keywords: Large Language Models, Knowledge Graphs, Retrieval-Augmented Generation, Explainable Artificial Intelligence, Domain-Specific Reasoning, Semantic Retrieval.

1. INTRODUCTION

Large language models (LLMs) have revolutionized natural language processing, showing impressive prowess in the areas of text generation, question answering, summarization, and reasoning. They have been adopted in various areas including, among others, healthcare, finance, law, engineering, and science research where knowledge is essential to get the right decisions in the right world. Standalone LLMs rely mainly on pre-trained

knowledge, however, and do not always receive information of a specialized nature or recent information, which results in hallucinations, incorrect facts, and unsound conclusions when answering complex domain-specific queries.

To enhance factual grounding, retrieval-Augmented Generation (RAG) has been proposed that retrieves relevant external documents before generating the response. Semantic vector retrieval, the most common approach for conventional RAG systems, has the ability to identify similar documents but may fall short in identifying explicit relationships between entities and concepts. Therefore, it may have limitations in addressing multi-hop reasoning and comprehensive evidence discovery. To complement vector-based retrieval, knowledge graphs representation of entities and their semantic relationships allows for the understanding and interpretation of the context in which these relationships appear.
While some actual graph-enhanced retrieval (GER) techniques have been suggested, most current ones lack reasoning transparency, lack enough evidence attribution, and fail to jointly consider graph traversal and semantic retrieval. These restrictions diminishes user trust, especially in knowledge-intensive systems for which explainability is crucial. Thus, we need a common model to integrate semantic retrieval with structured graph reasoning for deriving correct and evidence-based answers.

The study aims to present a Graph-Augmented Retrieval Framework for Explainable Domain-Specific Large Language Model Reasoning, combining semantic vector retrieval, knowledge graph traversal, evidence fusion, and generation of explanations to provide human-readable answers. The framework's goal is to make the retrieval more precise, the answers more accurate, the explanations more explainable, the facts more consistent and to minimize hallucination, while also ensuring acceptable response latency. This paper's contributions are a hybrid retrieval mechanism, an explainable reasoning pipeline, and an extensive evaluation via retrieval precision, answer accuracy, explainability score, rate of hallucination, and response latency. This paper has following sections: Related work is discussed in section 2, section 3 presents the proposed methodology, section 4 presents the results of the proposed methodology and section 5 concludes the paper.

2. LITERATURE REVIEW

Large Language Models (LLMs) have been shown to excel in knowledge-intensive tasks such as question answering, reasoning, text generation, and more, by learning rich linguistic and semantic representations from massive amounts of data [2]. This context-awareness has led to practical uses in medical areas, the financial and legal sectors, engineering and scientific research—areas that demand precise reasoning on specific areas of knowledge [3]. Pretrained language models are widely adapted by domain specific fine-tuning or using external knowledge during inference, (also referred to as retrieval) [12] to enhance performance in the specialized domain. Such adaptation methods boost the quality of the responses and minimize the need of relying on static, pre-trained knowledge [6].
Retrieval-Augmented Generation (RAG) has proven to be a powerful method for improving the factual correctness of LLMs by fetching external documents that are believed to contain relevant information prior to generating the response [1]. To locate semantically related information in large knowledge bases, there are some sparse, some dense and some hybrid retrieval approaches already in use in existing RAG systems [4]. While dense retrieval usually achieves higher semantic similarity compared to keyword-based approaches, the latter tends to miss explicit links between entities and concepts [1]. Consequently, traditional RAGs can have difficulties with multi-hop reasoning, understanding context and finding all relevant evidence for complex domain-specific queries [5].
Knowledge graphs (KGs) break these obstacles by modeling the domain knowledge as a graph of entities and semantic relations, allowing for a structured approach to reasoning with graph traversal [7]. Structured knowledge representations are further improved with knowledge graph construction and refinement [9]. Semantic dependency and connectivity between entities have also been enhanced by relational learning techniques in knowledge graph construction [11]. Recent research [10] has investigated using KG with LLMs for boosting reasoning accuracy, contextual comprehension, and generating responses with evidence. Furthermore, to increase transparency, there are explainable artificial intelligence techniques introduced such as attribution of evidence and providing reasoning paths and interpretable decision-making [13]. Such methods help users to be more confident of the knowledge that backs up the generation of a response [14].
While these progressions hold significant promise, current GraphRAG and explainable RAG solutions have several drawbacks, such as the poor level of explanation quality, the lack of evidence attribution, the limited

integration of the vector retrieval and graph reasoning modules, and the decreased effectiveness when using it for complex domain-specific reasoning tasks [8]. The aforementioned limitations call for a comprehensive retrieval system that integrates semantic vector retrieval, knowledge-graph traversal, evidence fusion and explainable reasoning. Hence, the study proposes to design a Graph-Augmented Retrieval Framework which provides greater precision in retrieval, accuracy in answers, explainability, and factual consistency in the given domain-specific LLM reasoning and reduces hallucinated responses.

3. METHODOLOGY

3.1 Overall Framework Architecture

The proposed framework involves a semantic vector retrieval, a graph reasoning, and a generation of explainable LLM, in a single architecture. It starts with the collection of domain documents, then text preprocessing, entity and relation extraction, construction of knowledge graph and generating the embedding of the documents. The embeddings reside in a vector database and the extracted entities; relationships reside in a graph database.

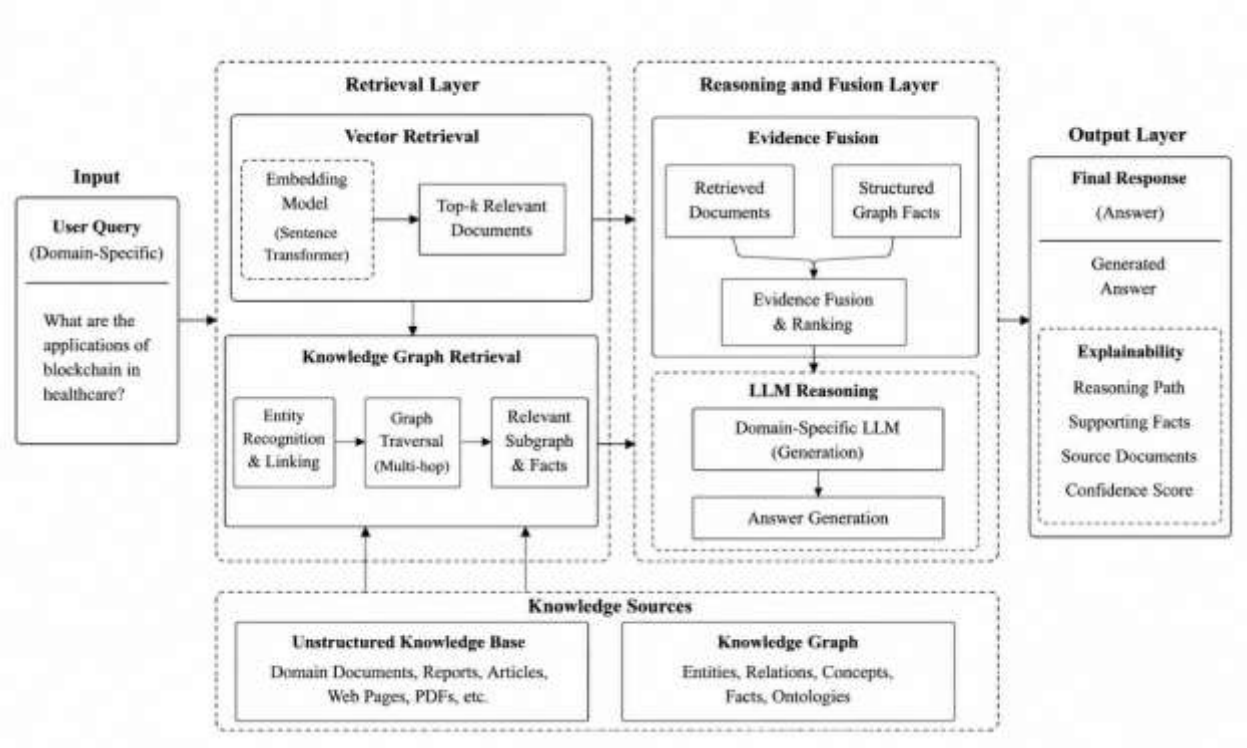

Figure 1. Overall Architecture of the Proposed Graph-Augmented Retrieval Framework

The framework will run parallel semantic vector retrieval and graph-based retrieval whenever a user issues a query. All the information retrieved is merged with an evidence-fusion module, which is then fed to an LLM reasoning module to generate the response. Lastly, the explanation-generation module gives the explanation path, reasons supporting the explanation, source identification, and returns the final answer. The integrated flow boosters the retrieval relevance, reasoning accuracy, factual consistency and explainability for applications in the domain.

3.2 Domain Knowledge Base and Graph Construction

The proposed scheme builds a domain knowledge base by gathering domain-specific documents from reports, article and technical manuals. These documents are first text cleaned and segmented, then entities are identified and relationships are extracted, and finally, entities are normalized. Receipt of the extracted entities and relationships to create the knowledge graph and document segments are converted into semantic embeddings. The knowledge graph is represented as

$$G = (V, E), \quad (1)$$

where Vrepresents domain entities and Erepresents semantic relationships between entities. The embeddings are stored in the FAISS vector database, whereas the graph is maintained in the Neo4j graph database. The knowledge base is periodically updated to incorporate new domain information. The configuration parameters used for graph construction, embedding generation, retrieval, and response generation are summarized in Table 1.

Table 1. Configuration Parameters of the Proposed Framework

Parameter	Description	Value
Document Chunk Size	Tokens in each document segment	512 tokens
Embedding Model	Semantic embedding model	Sentence Transformer
Embedding Dimension	Size of embedding vector	768
Vector Database	Storage for document embeddings	FAISS
Graph Database	Storage for entities and relationships	Neo4j
Similarity Metric	Vector similarity computation	Cosine Similarity
Top-k Retrieval	Number of retrieved document chunks	5
Graph Traversal Depth	Maximum traversal level	2 hops
Evidence Fusion Method	Combines vector and graph evidence	Weighted Ranking
Large Language Model	Reasoning and response generation	Llama 3
Response Temperature	Controls generation randomness	0.2
Maximum Response Length	Maximum generated output	512 tokens

The framework uses a document chunk size of 512 tokens, a Sentence Transformer embedding model that takes 768-dimensional embeddings, and Cosine Similarity for semantic retrieval. The FAISS vector database and Neo4j graph database store the vector and graph representations, respectively. The retrieval module retrieves top-5 relevant doc chunks and traverses graph up to 2 hops. Evidence is combined with the Weighted Ranking strategy, and Llama 3 is used to generate the response with a max length of 512 tokens and a temperature of 0.2.

3.3 Graph-Augmented Hybrid Retrieval

The proposed framework will use two retrieval paths: vector retrieval and graph retrieval, for each user query. In the vector retrieval path, the user query is converted to a semantic embedding and then compared to the document embeddings in the stored vector database to look for similar document chunks at the highest cosine similarity. The cosine similarity is computed as

$$S(q, d) = \frac{q \cdot d}{\|q\|\|d\|}, \quad (2)$$

where s(q,d) represents the cosine similarity score between the query and the document, q denotes the query embedding vector, d denotes the document embedding vector, $q \cdot d$ represents the dot product of the query and document vectors, and $\|q\|$ and $\|d\|$ represent the Euclidean (L2) norms of the query and document vectors, respectively. Simultaneously, the graph retrieval module extracts important entities from the query and matches them with the corresponding nodes in the domain knowledge graph. Controlled graph traversal retrieves related entities and semantic relationships to capture contextual information beyond vector retrieval. The retrieved evidence from both paths is then combined using

$$S_f = \alpha S_v + \beta S_g, \quad (3)$$

where $S_v$represents the vector-similarity score, $S_g$represents the graph-relevance score, αand βare the weighting parameters assigned to the vector retrieval and graph retrieval scores, respectively, and $S_f$denotes the final evidence score. The retrieved evidence is ranked according to S_f, and the highest-ranked evidence is supplied to the LLM reasoning module for generating an accurate, evidence-supported, and explainable response.

3.4 Explainable LLM Reasoning and Evaluation Setup

The retrieved evidence from the vector and graph retrieval modules are fed into a prompt for the LLM based on the evidence. The LLM is used to carry out the entity–relation reasoning, derive the reasoning path, assign source of the reasoning, and generate natural language explanations for increased transparency. The framework is tested with a domain specific corpus, Sentence Transformer embedding model, Neo4j graph database, and FAISS vector database, for embedding, retrieval, and graph structure building. Performance is compared to five baseline strategies: Standalone LLM, Conventional Vector-Based RAG, Knowledge-Graph-Only Retrieval, Existing GraphRAG and the Proposed Graph-Augmented Retrieval Framework. Five metrics are used for assessment: retrieval precision, accuracy of answers, explainability score, hallucination rate, and respone.

4. RESULTS AND DISCUSSION

4.1 Retrieval Performance Analysis

The retrieval performance of the proposed framework is analyzed by comparing retrieval precision of five methods. The proposed Graph-Augmented Retrieval Framework outperforms Existing GraphRAG (78.2%), Conventional Vector-based RAG (68.3%), Knowledge-Graph-Only Retrieval (61.4%) and Standalone LLM (52.6%) as shown in Figure 2. The results show the effectiveness of the use of knowledge-graph reasoning along with semantic vector retrieval.

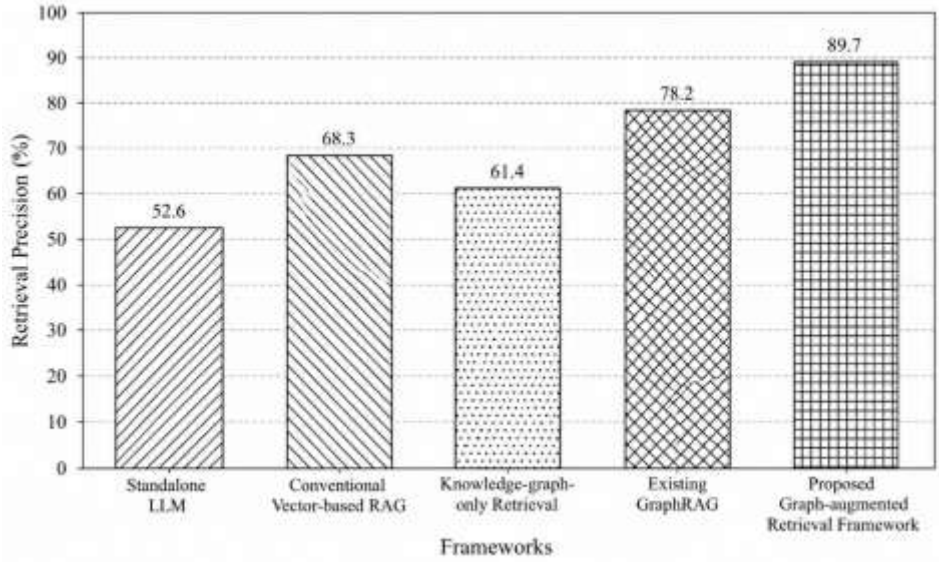

Figure 2. Retrieval Precision Comparison of Different Frameworks

The combined use of semantic similarity and graph relationships enables the retrieval of both directly relevant and indirectly related information while reducing irrelevant document retrieval. Thus, the proposed framework can more accurately provide evidence of domain-specific reasoning in various levels of query complexity.

4.2 Domain-Specific Answer Accuracy

The answer accuracy for the proposed framework is estimated through factual, relational and multi-step reasoning question types. The proposed framework is based on the various types of information such as semantic information from retrieved documents and structured information from the knowledge-graph, which helps it provide more accurate and context-aware responses compared with the baseline methods. Structured graph knowledge facilitates the LLM to recognize entity relationships and maintain the factual consistency when generating the answer. By combining unstructured textual evidence with structured graph knowledge, the proposed Graph-Augmented Retrieval Framework is able to answer more complex domain-specific queries with higher answer accuracy. The proposed method demonstrates superior performance across key metrics, including factual accuracy, relational accuracy, multi-step reasoning accuracy, and the number of reasoning errors and incomplete answers, compared to all other methods, such as Standalone LLM, Conventional Vector-Based RAG, Knowledge-Graph-Only Retrieval, and Existing GraphRAG.

4.3 Explainability and Evidence Analysis

The following aspects are analyzed: Quality of generated explanations, supporting evidence, reasoning paths, and source attribution, to evaluate the explainability of the proposed framework. The knowledge-Graph reasoning and semantic retrieval enable the framework to output a clear response by using explicit entities, relations, and documents to facilitate the generation of the response. This reasoning, based on evidence, not only helps to boost user confidence, but also allows them to check the generated answers. Unlike the baseline methods, the proposed Graph-Augmented Retrieval Framework can give more comprehensive reasoning paths and more relevant supporting evidence. The entity relationships are still evident in the reasoning process and make the sense of the retrieved knowledge clear for the user. Beyond that, human evaluation has shown that the outputted explanations are more informative, consistent and easier to understand. A typical explanation follows this pattern: Query → Retrieved Entity → Related Graph Nodes → Supporting documents → Generated Answer in order to provide transparent and explainable reasoning for the domain-specific question.

4.4 Hallucination and Response-Latency Analysis

The hallucination performance of the considered frameworks is compared in Figure 3. The Standalone LLM is particularly prone to hallucination with a rate of 25.6%, while Conventional Vector Based RAG rates at 16.8%, Knowledge-Graph-Only Retrieval at 13.2%, and Existing GraphRAG at 8.7%. The Proposed Graph-Augmented Retrieval Framework (PGRF) reduces the number of unsupported claims with evidence grounding and structured graph reasoning, achieving the lowest hallucination rate of 4.1%.

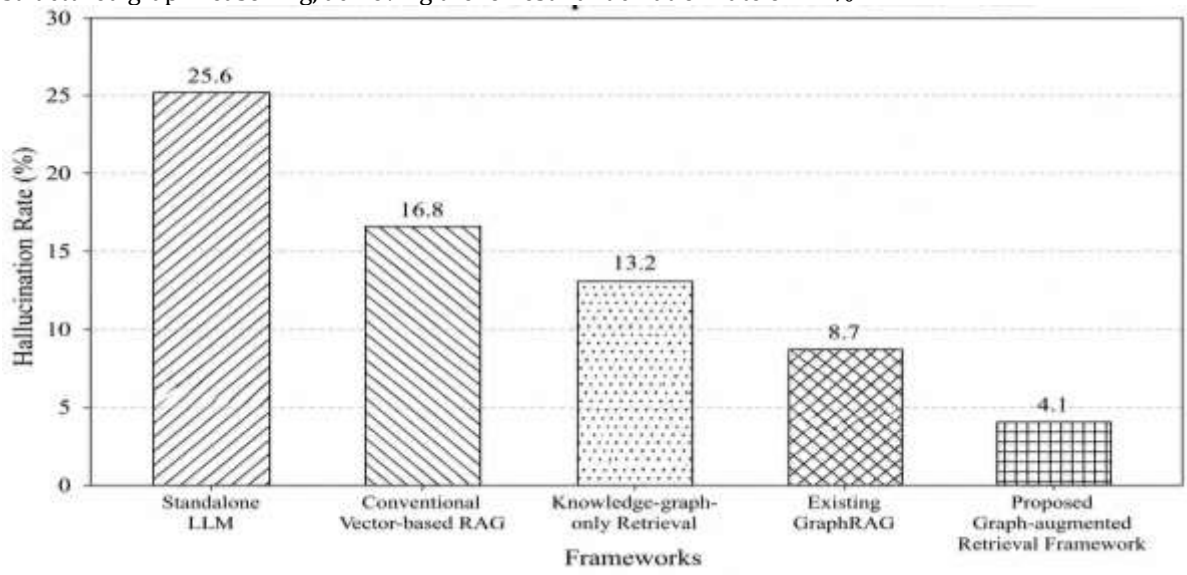

Figure 3. Hallucination Rate Comparison of Different Frameworks

The increase in response latency with graph traversal and evidence fusion is not significant, and the benefit of having a more factually reliable and accurate reasoning account is still valid for the moderate cost. The proposed framework strikes an appropriate balance between response time and the accuracy and explainability of the answer provided, thus being suitable to different domain-specific applications that require accurate and explainable outputs even though the response time isn't the major concern there.

4.5 Overall Comparative and Ablation Analysis

The overall performance of the analysed reasoning frameworks is summarised in Table 2, which shows the precision of retrieval results, accuracy of responses, explainability score, rate of hallucinations and latency of responses. With 89.7% retrieval precision, 95.3% answer accuracy, an explainability score of 9.6 and the fewest hallucination errors (4.1%), the proposed Graph-Augmented Retrieval Framework achieves the best performance, while also having a 492ms response latency.

Table 2. Overall Performance Comparison of Different Reasoning Frameworks

Framework	Retrieval Precision (%)	Answer Accuracy (%)	Explainability Score (1–10)	Hallucination Rate (%)	Response Latency (ms)
Standalone LLM	52.6	71.8	5.6	25.6	320

Conventional Vector-Based RAG	68.3	82.7	7.2	16.8	385
Knowledge-Graph-Only Retrieval	61.4	79.5	8.1	13.2	410
Existing GraphRAG	78.2	89.4	8.8	8.7	465
Proposed Graph-Augmented Retrieval Framework	89.7	95.3	9.6	4.1	492

The ablation study is conducted while discarding the explanation module, evidence fusion, vector retrieval, and graph retrieval. From the results, one can see that both components have a positive effect on the quality of retrieval, reasoning accuracy, explainability and reduction of hallucination. The whole proposed configuration is the one that performs best in all tested configurations.

5. CONCLUSION

In this paper, a Graph-Augmented Retrieval Framework for Explainable Domain-Specific Large Language Model Reasoning was proposed, which combines the semantic vector retrieval method, knowledge graph, and large language models, and establishes an integrated retrieval and reasoning pipeline to achieve explainable reasoning with domain-specific LLMs. The proposed framework is able to effectively leverage unstructured textual evidence and structured graph knowledge to achieve better precisions in the domain-specific retrieval, better answer accuracy, and better explainability. The evidence-based reasoning process greatly reduces the number of unsupported and hallucinated responses and also creates clear reasoning paths with supporting evidence to drive the generated response and makes it easier to interpret and understand the answers generated than conventional retrieval-augmented generation systems. Experimental results show that the proposed approach consistently leads to better performance compared with standalone LLMs, the conventional vector-based RAG approach, the knowledge-graph-only retrieval approach, and existing GraphRAG approaches, on all the evaluation metrics. The cost of constructing the graph and traversing it also adds to the computational complexity, but it is justified by the benefits of enhancing factual reliability and reasoning quality, even if the performance of the graph construction and graph traversal algorithm are affected by the quality of the entity and relation extraction. Further enhancements of the documents, knowledge graphs, and reasoning capabilities will focus on multimodal document integration, dynamic updates for the knowledge graphs, larger domain-specific corpora, and adaptive retrieval strategies to boost scalability, robustness, and reasoning performance.

REFERENCES

1. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems, 33, 9459-9474.
2. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.
3. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.
4. Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M. (2020, November). Retrieval augmented language model pre-training. In International conference on machine learning (pp. 3929-3938). PMLR.
5. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of hallucination in natural language generation. ACM computing surveys, 55(12), 1-38.
6. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.
7. Ji, S., Pan, S., Cambria, E., Marttinen, P., & Yu, P. S. (2021). A survey on knowledge graphs: Representation, acquisition, and applications. IEEE transactions on neural networks and learning systems, 33(2), 494-514.
8. Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., Melo, G. D., Gutierrez, C., ... & Zimmermann, A. (2021). Knowledge graphs. ACM Computing Surveys (Csur), 54(4), 1-37.
9. Paulheim, H. (2016). Knowledge graph refinement: A survey of approaches and evaluation methods. Semantic web, 8(3), 489-508.

10. Ji, S., Pan, S., Cambria, E., Marttinen, P., & Yu, P. S. (2021). A survey on knowledge graphs: Representation, acquisition, and applications. IEEE transactions on neural networks and learning systems, 33(2), 494-514.
11. Nickel, M., Murphy, K., Tresp, V., & Gabrilovich, E. (2015). A review of relational machine learning for knowledge graphs. Proceedings of the IEEE, 104(1), 11-33.
12. Wang, R., Tang, D., Duan, N., Wei, Z., Huang, X. J., Ji, J., ... & Zhou, M. (2021, August). K-adapter: Infusing knowledge into pre-trained models with adapters. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021 (pp. 1405-1418).
13. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.
14. Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE access, 6, 52138-52160.