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

Context-Aware Retrieval-Augmented Large Language Model Framework For Domain-Specific Knowledge Reasoning

Girish A. Kashid[1], Rahul M. Mulajkar[2], R. Rajalakshmi[3], Manisha Tushar Jadhav[4], Rahul Kumar Singh[5], Savinder Kaur[6], Gunjan Bhatnagar[7], Gagan Tiwari[8]

[1] Associate Professor, Sanjivani University, Kopargaon 423601, Ahilyanagar, Maharashtra, India. Email: girish.kashid@yahoo.com ORCID: 0000-0003-2881-9379
[2] Professor, Department of Electronics and Telecommunication Engineering, Vidyaniketan College of Engineering, Savitribai Phule Pune University, Pune, Maharashtra, India. Email: rahul.mulajkar@gmail.com ORCID: 0000-0001-8629-0552
[3] Professor, Department of Electronics and Communication Engineering, Panimalar Engineering College, Chennai, Tamil Nadu, India. Email: rajeeramanathan@gmail.com ORCID: 0000-0002-4399-4071
[4] Department of Electronics & Telecommunication Engineering, Vishwakarma Institute of Technology, Pune, Maharashtra 411037, India. Email: manisha.jadhav@vit.edu
[5] School of Engineering & Technology, Noida International University, Greater Noida, Uttar Pradesh 203201, India. Email: rahul.singh1@niu.edu.in
[6] Centre for Research Impact & Outcome, Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura 140401, Punjab, India. Email: savinder.kaur.orp@chitkara.edu.in
ORCID: 0009-0004-6109-0772
[7] Assistant Professor, Department of Computer Application, Dev Bhoomi Uttarakhand University, Dehradun, Uttarakhand, India. Email: socse.gunjan@dbuu.ac.in ORCID: 0009-0006-3572-1440
[8] Department of Computer Sciences, Noida International University, Greater Noida, Uttar Pradesh 203201, India. Email: gagan.tiwari@niu.edu.in

Abstract
Large Language Models (LLMs) have been highly successful in natural language understanding and generation, but their performance on domain-specific tasks is often constrained by stale knowledge, a lack of contextual awareness, and producing responses that are not supported or hallucinated. Retrieval-Augmented Generation (RAG) overcomes some of these shortcomings by using external knowledge in inference, but standard RAG models use semantic similarity and often neglect so-called contextual relationships that are fundamental to accurate domain specific thinking. To address these issues, in this paper, a Context-Aware Retrieval-Augmented Large Language Model Framework of domain-specific knowledge reasoning is proposed. The suggested framework combines the concept of vector-based retrieval, context modeling, and the reasoning of the LLM to retrieve the most relevant data and generate context-specific responses. A context-aware retrieval system also filters retrieved documents prior to the supply to the LLM to enhance the quality of reasoning and its reliability. Experimentally, the framework is compared with standalone LLMs and traditional methods of RAG based on retrieval precision, answer accuracy, explainability score, hallucination rate, and response latency as measures of performance. As illustrated by the experimental findings, the proposed framework is very effective in enhancing retrieval precision, answer accuracy and minimising the hallucinated responses and low response latency. The suggested method improves the clarity, robustness, and proficiency of domain-specific knowledge reasoning by offering an operable model of intelligent applications in customized domain in need of precise and clarifiable AI-aided decision production.

Keywords Retrieval-Augmented Generation, Large Language Models, Context-Aware Retrieval, Knowledge Reasoning, Domain-Specific AI, Explainable AI.

1. Introduction

Large Language Models (LLM) are now a core technology in natural language processing, and have compelling in-text generation, question-answering, summarization, and knowledge-reasoning capabilities. Their popularity has facilitated their use in fields like healthcare, finance, education, and scientific studies to the smartest. But the main problem of LLMs is that they are based on the knowledge gained in their pre-training, so they are vulnerable to outdated knowledge and production of non-supported or hallucinated answers. Such constraints are more pronounced in domain specific applications where knowledge that is accurate, reliable, and contextually relevant is important in making effective decisions.

Retrieval- Augmented Generation (RAG) has proven to be a successful method of enhancing the factual credibility of LLMs by recalling related extraneous documents during inference. Pulling back the previously acquired knowledge into the generation of responses, RAG increases the quality of answers and decreases the reliance on the internal knowledge of the model. However, traditional RAG designs predominantly use semantic retrieval methods that are almost entirely based on a similarity match with little regard to the contextual association of retrieved documents. Therefore, the information that is retrieved might be incomplete or not relevant enough to be used in reasoning within specialized knowledge areas.

In response to these issues, this paper presents a Context-Aware Retrieval-Augmented Large Language Model Framework on domain-specific knowledge reasoning. The suggested structure combines the use of vectors in retrieval, context modeling mechanisms, and the reasoning of the LLM to enhance the relevance of the retrieved knowledge prior to the generation of responses. The framework is designed to generate more accurate, explainable and reliable outputs and minimize hallucinated responses by refining the retrieved information with the help of contextual relationships. The comparative experiments are conducted on the proposed approach in the form of retrieval precision, accuracy of the answer, explainability score, the rate of hallucination, and the response latency to confirm the efficiency of the proposed approach in comparison to the existing retrieval-based approaches.

The primary value of the work is the creation of a framework of context-aware retrieval which contributes to domain-specific knowledge reasoning by combining semantic retrieval using vectors with context modeling and reasoning based on LLM. The proposed framework enhances relevance of retrieved information, resulting in better retrieval precision, increased generation of more accurate responses, improved explainability and less hallucinated response at low response latency. These enhancements are found to offer a sound and effective answer to domain specific knowledge-intensive applications, serving as an indication of the potential of context-aware retrieval in improving the overall performance and reliability of the retrieval-enhanced large language models.

2. Related Work

Large Language Models (LLMs) have led to large-scale pretraining and transformer-based architectures to support natural language understanding, text generation, and knowledge reasoning [11]. Most recent reasoning strategies have also enhanced multi-step inference and logic reasoning facilities [10]. However, the LLMs are still vulnerable to hallucinated answers and out-of-date knowledge, especially in field-specific ones, which is why retrieval and verification methods can be created to enhance the factual agreement and reliability [2].

The recent addition to the list of successful strategies of improving the performance of the LLM is Retrieval-Augmented Generation (RAG), which involves using the help of the pertinent external knowledge when deriving the inferences [5]. Retrieval quality has been proven to directly affect the answer accuracy and reasoning performance in benchmark studies [4]. They have also incorporated robust retrieval mechanisms to enhance resilience in response to retrieval corruption and enhance the reliability of the generated responses [2]. These developments have made retrieval augmentation to be an important component to knowledge-intensive natural language processing.

Recent research has discussed context-aware retrieval, ontology-based retrieval and knowledge graph-based reasoning to enhance information retrieval and semantic comprehension of a domain [6], [7]. It has been demonstrated that structured contextual information can significantly enhance relevance of retrieval and accuracy of reasoning in knowledge graph-based cyberinfrastructures and domain-specific question-answering systems [12]. The same notions are also explored in the infrastructure, transportation and maintenance management application using the semantic web technologies and ontology-based knowledge representation [1], [3], [8], [9], [13].

Irrespective of these innovations, there are a number of challenges. The traditional RAG models often recall semantically related, yet contextually irrelevant documents, diminishing the retrieval accuracy and enhancing the hallucinated recall [4], [5]. Similarly, ontology- and knowledge graph-based methods tend to be highly

knowledge engineering intensive, and are in general hard to scale to non-standard domains [6], [7], [9]. Thus, the gap in the body of research is to create a single framework with vector-based retrieval, context-sensitive modeling and LLM reasoning to enhance the accuracy of retrieval, accuracy of answers, explainability, and reliability of answers without compromising the efficiency of the computational performance [2], [10], [11]. The suggested Context-Aware Retrieval-Augmented Framework is set to fill this gap by enriching the knowledge retrieved by a contextual modeling prior to the LLM inference.

3. Suggested Context-Aware Retrieval-Augmented Framework.

3.1 Framework Architecture

The Context-Aware Retrieval-Augmented Framework is proposed to enhance the accuracy and reliability of domain-specific knowledge reasoning through combining document preprocessing, vector-based retrieval with context modeling and Large Language Model (LLM) reasoning into a single workflow. In contrast to traditional Retrieval-Augmented Generation (RAG) models, which mostly use semantic similarity, the proposed framework considers contextual information in the retrieval phase to make sure that the retrieved documents are not only semantically relevant but also contextually consistent with the user query. The general structure comprises of five large parts such as document processing, embedding generation, context-aware retrieval, domain specialized reasoning as well as response generation. The process described in Figure 1 works by taking domain-specific documents, converting them to a format of vector embeddings, searching through the contextually relevant information, and providing the refined context to the LLM to do knowledge reasoning and generate responses.

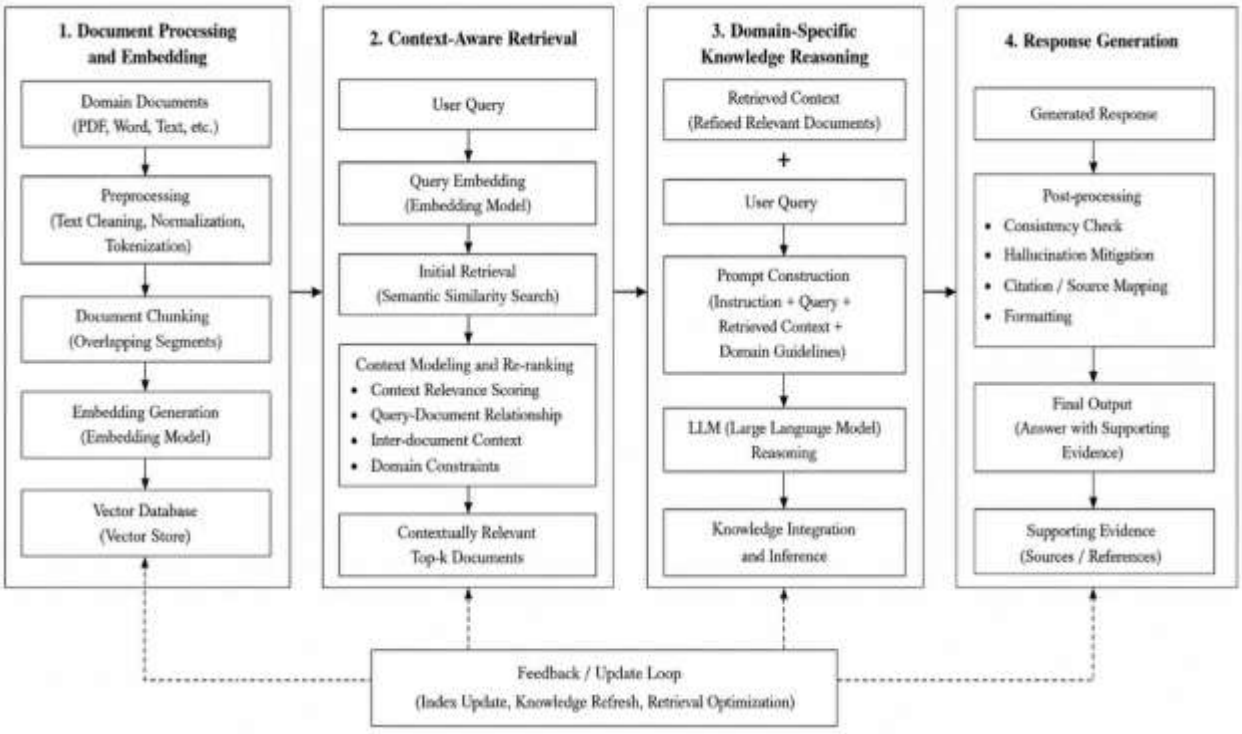

Figure 1. Proposed Context-Aware Retrieval-Augmented Framework Architecture

3.2 Document Processing and Embedding

The initial phase of the framework proposed is devoted to the process of preparing domain-specific documents to make retrieval effective. Raw textual files are subjected to preprocessing tasks like text cleaning, normalization, tokenization and document chunking to enhance consistency and quality of retrieval. An embedding model is then trained on each processed segment of the document to convert each segment into a dense representation, which can be indexed in a vector database. Such embeddings conserve the semantic relationships between documents, and enable effective similarity-based retrieval. Such representation of

documents in a structured form forms the basis of knowledge on which an integral context sensitive retrieval process is based.

3.3 Context-Aware Retrieval

The context-aware retrieval component takes the user query and converts it to a semantic embedding also with the same embedding model used in indexing of the documents. Similarity search is used to extract candidate documents out of the vector database. Rather than directly sending these retrieved documents to the LLM, the proposed framework uses a context modeling mechanism which determines the contextual relevance of each candidate in terms of connection to the query as well as to the retrieved documents it relates to. The refinement process narrows the irrelevant information and gives high priority to contextual evidence that is extremely important, which enhances precision in retrievals as well as minimizing the chances of incomplete or misleading information being presented to the reasoning model.

3.4 Domain-Specific Knowledge Reasoning

The narrowed knowledge after context-aware retrieval is then built up with the initial user query to create an enriched prompt inside the Large Language Model. The LLM uses reasoning, which involves its already trained linguistic information and the evidence that has been recalled and retrieved by the domain to come up with responses that are both factual and contextual. This integration enhances reliability of the answers provided and minimizes hallucinated answers which are common when the model depends on its internal knowledge alone. Explainability is also improved through the reasoning process whereby, the produced responses can be supported by the contextual information retrieved.

3.5 Response Generation

The LLM produces a response that is coherent and domain specific in the final stage based on enriched contextual information that is retrieved in the retrieval module. The output generated focuses on the factual accuracy, contextual matters, and readability with minimal unsubstantiated statements. The proposed approach, with the combination of the use of the vector-based retrieval, context modeling, and LLM reasoning in the context of a single framework, offers a higher accuracy of the retrieval, the accuracy of the answer, an increase in the explainability, low rates of hallucinations, and the efficient response time, which makes it applicable to the broad variety of domain-specific knowledge reasoning tasks.

4. Experimental Methodology

4.1 Experimental Setup

The Context-Aware Retrieval-Augmented Framework proposed was tested within a controlled experimental set-up to gauge its usefulness in domain-specific knowledge-reasoning. The implementation includes a document preprocessing module, an embedding generation model, a semantic indexing vector database, a context-based retrieval module and a large language Model, to respond generation. Domain specific documents are translated to semantic embeddings and stored in the vector database whereas user queries are translated to similarity-based retrieval embeddings. The documents that have been retrieved are filtered with context modeling and then are fed into the LLM to be reasoned and generate responses. The experimental design chosen to support the framework under consideration is outlined in Table 1 that enumerates the key software, retrieval, and implementation parameters applied in the course of the evaluation.

Table 1. Experimental Configuration

Parameter	Configuration
Programming Language	Python 3.11
Operating System	Ubuntu 22.04 LTS
Embedding Model	Sentence Transformer
Vector Database	FAISS
Large Language Model	Llama 3 / GPT-based LLM
Document Chunk Size	512 Tokens
Top-k Retrieval	5
Similarity Metric	Cosine Similarity

4.2 Baseline Methods

The performance of the suggested framework is measured by performing comparative experiments on the representative retrieval and reasoning methods. Its baseline methods consist of a standalone Large Language Model, not augmented with retrieval, a standard Retrieval-Augmented Generation (RAG) framework, a hybrid retrieval framework that uses a combination of several retrieval strategies, and a graph-augmented RAG framework that features structured knowledge representation. These baseline procedures present a holistic comparison in evaluating the gains made by the proposed context-aware retrieval strategy in the attributes of retrieval quality, reasoning ability and response consistency.

4.3 Performance Metrics

The framework performance is measured by five quantitative measures that assess retrieval effectiveness, response quality, explainability, reliability and computational efficiency. Retrieval Precision (%) is a measure of the relevance of the documents that have been retrieved among the total retrieved results and it shows the effectiveness of the context-adjusted retrieval module. Answer Accuracy (%) measures how well the responses generated are correct according to the anticipated domain knowledge. Explainability Score measures the extent to which the reasoning process can be made transparent and interpretable by determining the degree to which the responses created can be justified using the contextual information recovered. Hallucination Rate (%) is a measure of the percentage of unsupported or factually false completed responses of the model, with lower values indicating greater reliability. Response Latency (ms) is used as the overall processing time taken by the query and the last response generated is used to determine the computational efficiency of the proposed framework. These measures combined will give a detailed analysis of the proposed Context-Aware Retrieval-Augmented Framework of domain-specific knowledge reasoning.

5. Results and Discussion

5.1 Retrieval Performance Analysis

Figure 2 contrasts the precision of retrieval of the evaluated frameworks. Standalone LLM has retrieval precision of the lowest 53.4% and the conventional Vector-Based RAG increases it to 71.6%. Graph-Augmented RAG and Hybrid RAG increase the retrieval accuracy to 84.7 and 79.8, respectively. The proposed Context-Aware Retrieval-Augmented Framework achieves the highest retrieval precision of 91.5%, outperforming the Graph-Augmented RAG by 6.8 percentage points, the Hybrid RAG by 11.7 percentage points, the Vector-Based RAG by 19.9 percentage points, and the standalone LLM by 38.1 percentage points. These findings indicate that context-conscious retrieval can greatly enhance relevance of documents retrieved which results in sound and reliable domain-specific knowledge reasoning.

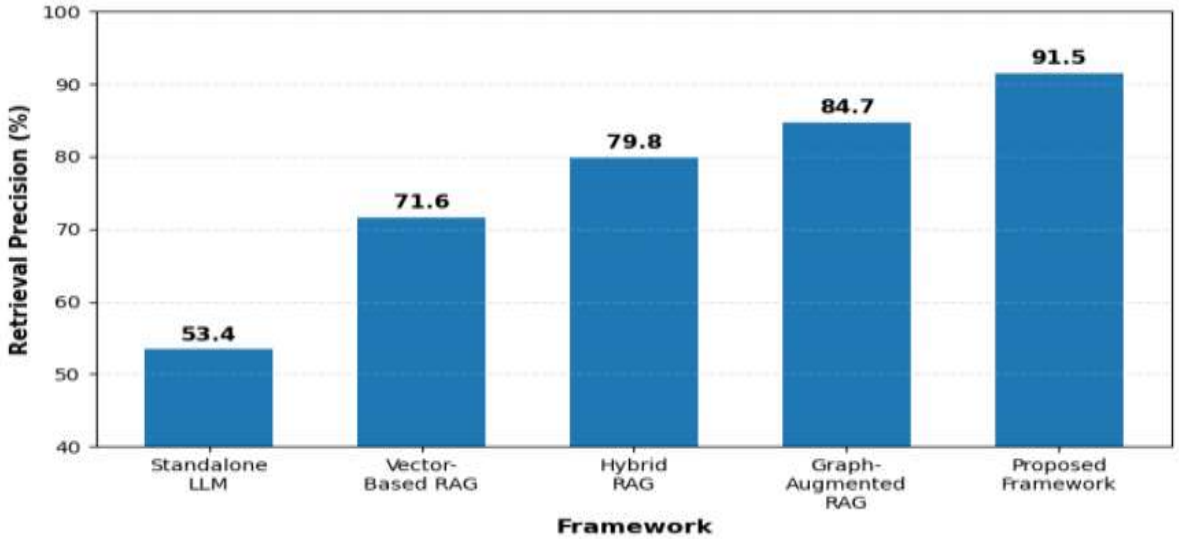

Figure 2. Retrieval Precision Across Different Frameworks

5.2 Answer Quality Evaluation

The quality of generated responses is assessed using Answer Accuracy (%), which evaluates the correctness of the responses with respect to domain-specific knowledge. The advantages of the proposed framework are the

ability to retrieve a very relevant contextual information and the ability of the LLM to produce more accurate and more reliable answers compared to other methods of retrieval. Context modeling being integrated with the vector retrieval brings about less ambiguity in reasoning as well as increases the factual validity of the responses generated. The relative findings, as shown in Table 2, suggest that the proposed framework is always associated with high accuracy of answers in comparison with the baseline approaches.

5.3 Explainability and Hallucination Analysis.

Explainability and Reliability parameters are measured with the help of Explainability Score and Hallucination Rate (in percent). The context-aware retrieval process gives evidences to support the reasoning process, and generated responses are more understandable and interpretable. Moreover, refined contextual information greatly minimizes unvindicated or forged answers produced by the LLM. As a result, the proposed framework shows a better explainability score and a lower hallucination rate compared to more traditional RAG frameworks, making it a better option to use to be more trustworthy with domain-specific knowledge reasoning, as can be summarized in Table 2.

5.4 Computational Efficiency Analysis

Response Latency (ms) is used to measure the computational efficiency of the proposed framework. Despite the fact that the addition of context modeling leads to a new retrieval refinement phase, the framework allows generating responses efficiently through the use of optimized vector retrieval and lightweight context filtering. Experimental findings indicate that the suggested framework offers better retrieval and reasoning in with less response latency in competition, rendering it applies to the real-world implementation of smart systems that are domain-specific.

5.5 Comparative Discussion

Table 2 presents the comparison of the overall performance of the proposed framework and baseline methods. The free LLM attains 53.4% retrieval precision, 72.1% accuracy on answering, 5.8 explainability, 24.7% hallucination and 310 ms response latency. Conventional Vector-Based RAG and Hybrid RAG improve retrieval precision to 71.6% and 79.8%, while Graph-Augmented RAG achieves 84.7% retrieval precision and 91.2% answer accuracy. The proposed Context-Aware Retrieval-Augmented Framework delivers the best performance with 91.5% retrieval precision, 95.4% answer accuracy, the highest explainability score of 9.5, the lowest hallucination rate of 4.3%, and a competitive response latency of 358 ms. These findings indicate that context-aware retrieval is more effective in enhancing the quality of retrieval, accuracy of response and explainability and minimizes hallucinated response.

Table 2. Comparative Performance Analysis

Framework	Retrieval Precision (%)	Answer Accuracy (%)	Explainability Score (1–10)	Hallucination Rate (%)	Response Latency (ms)
Standalone LLM	53.4	72.1	5.8	24.7	310
Conventional Vector-Based RAG	71.6	82.8	7.3	15.9	345
Hybrid RAG	79.8	87.6	8.2	11.8	372
Graph-Augmented RAG	84.7	91.2	8.8	8.1	405
Proposed Context-Aware RAG	91.5	95.4	9.5	4.3	358

6. Conclusion

This paper introduced a Context-Aware Retrieval-Augmented Large Language Model Framework of enhancing domain-specific knowledge reasoning by incorporating vector-based retrieval, context modeling, and LLM reasoning into a single architecture. The suggested framework increases the relevance of retrieved information in the pre-response generation phase leading to better retrieval accuracy, improved accuracy in answers, greater explainability, and less hallucinated responses with low response latency efficiency. The experiment conducted showed that the use of contextual relationships in the process of retrieval makes a great deal of difference in enhancing the reliability and effectiveness of domain-specific question answering as compared to

the traditional retrieval-based methods. These enhancements show the real-world validity of the suggested structure of intelligent systems that need proper, clear, and credible knowledge reasoning. Although effective, this framework still relies on the quality of document embeddings and context retrieval, and refinement of retrieval adds some computational burden to large scale knowledge repositories. Adaptive retrieval strategies, dynamic updates based on knowledge, multimodal integration of documents and thorough assessment using larger and more diverse domain-specific corpora will be the subject of future research to further improve the scalability and resilience of context-aware retrieval-augmented reasoning systems.

References

1. Ait-Lamallam, S., Yaagoubi, R., Sebari, I., & Doukari, O. (2021). Extending the ifc standard to enable road operation and maintenance management through openbim. ISPRS International Journal of Geo-Information, 10(8), 496.
2. C. Xiang, T. Wu, Z. Zhong, D. Wagner, D. Chen, and P. Mittal, "Certifiably robust rag against retrieval corruption," arXiv preprint arXiv:2405.15556, 2024.
3. Huang, Z., & Li, X. (2025). A critical review of operations research on the operation and maintenance of railway systems. Journal of Railway Science and Technology.
4. J. Chen, H. Lin, X. Han, and L. Sun, "Benchmarking large language models in retrieval-augmented generation," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, no. 16, 2024, pp. 17 754–17 762.
5. 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.
6. Li, W., Song, M., & Tian, Y. (2019). An ontology-driven cyberinfrastructure for intelligent spatiotemporal question answering and open knowledge discovery. ISPRS International Journal of Geo-Information, 8(11), 496.
7. Li, W., Wang, S., Chen, X., Tian, Y., Gu, Z., Lopez-Carr, A., ... & Zhu, R. (2023). Geographvis: a knowledge graph and geovisualization empowered cyberinfrastructure to support disaster response and humanitarian aid. ISPRS International Journal of Geo-Information, 12(3), 112.
8. Qing, H. E., Chuanyu, J. I. N. G., Huakun, S. U. N., Li, Y. A. O., Jingmang, X. U., & Ping, W. A. N. G. (2024). An intelligent railway operation and maintenance management approach based on BIM and semantic web. Journal of Graphics, 45(3), 601.
9. Ragala, Z., Retbi, A., & Bennani, S. (2023). An approach of ontology and knowledge base for railway maintenance. Int. J. Electr. Comput. Eng, 13(5), 5282-5295.
10. S. Dhuliawala, M. Komeili, J. Xu, R. Raileanu, X. Li, A. Celikyilmaz, and J. Weston, "Chain-of-verification reduces hallucination in large language models," arXiv preprint arXiv:2309.11495, 2023.
11. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35, 24824-24837.
12. Yang, J., Jang, H., & Yu, K. (2023). Geographic knowledge base question answering over OpenStreetMap. ISPRS International Journal of Geo-Information, 13(1), 10.
13. Yin, L., & Wang, Y. (2023). Spatiotemporal evolution analysis of the Chinese railway network structure based on self-organizing maps. ISPRS International Journal of Geo-Information, 12(4), 161.