Grenze International Journal of Engineering and Technology, June Issue

Graph-based Evidence Verification and Explainable RAG Framework for Space Missions

Sameera P Simha[1] and Meena Kumara K.S[2]
[1]PG Scholar, Department of Computer Science and Engineering, Presidency University, Bengaluru, Karnataka, India
Email: sameera.2024DSC0002@presidencyuniversity.in
[2]Assistant Professor, Department of Computer Science and Engineering, Presidency University, Bengaluru, Karnataka, India
Email: meenakumari.ks@presidencyuniversity.in

***Abstract*— This paper proposes a fresh approach to checking facts uses graphs, built into a clear step-by-step process meant for trustworthy help in space missions. Instead of just pulling data, it finds official manuals, checks details through linked logic, then answers with reasons shown. What sets it apart is how it says no when unsure - no guesses made. Running without internet access, it relies on a freely available tuned language model that follows directions well. Tests show fewer mistakes, less made-up output, better visibility into decisions than older methods.**
A fresh approach called Retrieval-Augmented Generation pipelines now helps LLMs tap into outside information. Still, current versions tend to invent details, offer no clear reasoning trail, nor allow checks on accuracy.
Because of these gaps - like unverified outputs and opaque logic - they struggle where trust matters most. Think spacecraft operations, where mistakes aren't an option. This paper proposes a method to work offline critical missions.

***Index Terms*— LLMs, RAG Pipelines, Explainable RAG, Offline systems, Space Missions.**

I. INTRODUCTION

Explainable Indian space missions create and use extensive collections of technical documents and mission records and design documents and user guides and scientific research material which they use throughout their entire mission execution process that starts with mission design and continues through system installation and mission launch and active mission activities and ends with post-mission evaluation. The accurate information from these different sources needs to be accessible because it supports mission planning and anomaly resolution and decision-making processes.
Organizations want to use artificial intelligence systems because their missions become more complex and they need to use autonomous systems which require the development of new artificial intelligence methods to help engineers and scientists with their essential mission information analysis tasks. LLMs provide advanced capabilities which enable computers to understand human language and answer specialized technical inquiries. The space mission environment becomes dangerous because the system generates misleading information which appears to be correct but contains significant errors because even the smallest mistakes in space missions can create large operational problems.

Grenze ID: 01.GIJET.12.2.43_26

II. Related Work

A. Preliminaries

Retrieval-Augmented Generation introduces a new way for large language models to access external knowledge sources which solves the main issue that limits standalone neural language models because they depend on fixed knowledge that their training process has established. The operational system of traditional large language models generates text which maintains coherence and contextual accuracy but functions as a closed system because it restricts knowledge to what existed at the time of their training. The architectural constraint results in three major problems which include systems that cannot retrieve new information and systems that produce incorrect specific domain knowledge and systems that create convincing yet false output through hallucination. RAG operates as a hybrid system which connects parametric and non-parametric memory systems through its process of retrieving relevant documents from an external corpus before using that information to generate new content. The basic design of RAG includes two main elements which function together as a unified system because it contains both a retriever and a generator. The retriever uses dense passage retrieval techniques and sparse lexical matching methods which include BM25 to find relevant document segments from its extensive document collection. The system transforms input queries into vector formats through neural encoders which then calculate similarity scores using existing document chunk embeddings stored in a vector database.

B. Literature Survey

Arora et al. [1] developed a three-layer system which improves reasoning abilities of large language models through its structured semantic reasoning paths system that prevents hallucination errors. The framework uses knowledge graph structures together with reasoning-based generation systems while its factual validation component controls semantic paths during inference operations. Wang et al. [2] developed a knowledge graph-guided framework specifically targeting faithful and explainable reasoning in large language models. Their approach uses structured knowledge integration to improve reasoning capabilities while they use knowledge graph structures to perform explicit factual validation, which reduces hallucination issues. The research establishes a theoretical framework which combines symbolic knowledge systems with neural generation technologies, but it does not address practical use cases for environments with limited resources. Bensch et al. [4] created CORE, which functions as a modular Intelligent Personal Assistant that helps astronauts with spaceflight procedural tasks by using retrieval-augmented generation and knowledge graph systems and augmented reality technologies. The authors incorporate Generative Pre-trained Transformers (GPT) such as GPT-4 for initial training purposes and Llama 3.1 for an offline capable system, as connectivity remains a key challenge in space missions. The authors use augmented reality signals in order to preserve the cognitive functions of astronauts through eliminating the switching between their operational tasks and checklist-based routines. Current techniques fail to provide sufficient support due to rigidity of rule-based approaches and excessive hardware required by leading proprietary models whereas open-source solutions provide the ability to function without internet connectivity. GraphRAG [7] is a method proposed by Diamantini et al. focused specifically on explainability of dataset discovery techniques. In particular, the authors develop a framework that leverages the advantages of knowledge graphs for query enrichment, dataset ranking, and explanation reporting. In addition, the proposed solution utilizes a combination of LLMs to transform queries and knowledge graphs to enrich semantic features while implementing preference-aware ranking approaches. It exhibits very good explainability capabilities through ensuring transparency in its discovery mechanism and accommodating user preferences. One of the main challenges faced during the implementation include complexity of pipeline design as well as dependency on the completeness and quality of knowledge graphs. Hua et al. [8] developed layer-wise relevance propagation methods to detect hallucinations in retrieval-augmented generation systems. The requirement for model internal access and computational expense for large-scale language models represent deployment limitations. The work provides valuable foundations for verification and explainability in safety-critical RAG applications. But these research works lack the basic requirements that is reducing the hallucination rates and increasing the transparency of the model. Though some techniques have initially changed the traditional dimension of RAG pipelines, they do not contribute towards reducing hallucination.

III. Research Gaps Identified

Retrieval-Augmented Generation systems have achieved major progress yet their current shortcomings make them unfit for use in high-stakes situations which include space exploration missions. Existing RAG systems continue to generate factually incorrect or ungrounded information even when provided with retrieved context, as current methods do not fully eliminate the risk of fabricated content, particularly when dealing with ambiguous or incomplete source documents. The standard RAG pipelines function on the assumption that retrieved documents provide both correct information and complete details required for query responses while the system lacks methods to confirm whether retrieved proof validates the produced answer or if different sources concur on vital information, which results in current systems failing to tackle cross-document validation challenges. Existing RAG systems cannot detect when their retrieved sources show opposite information because they only use selective evidence that supports their claims while omitting proof that contradicts them or they create unclear answers by merging different pieces of opposing information because current systems lack tools to automatically find inconsistencies between multiple sources.

IV. METHODOLOGY

A. Proposed System Architecture

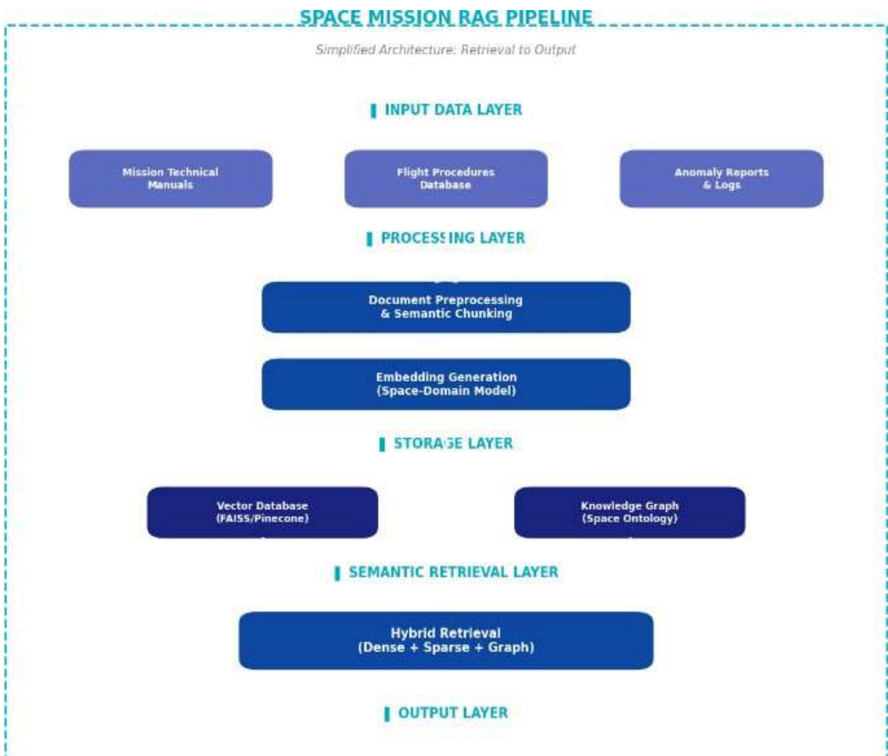

Fig. 1. Proposed System Architecture

Fig. 1 displays a simple architecture diagram displays a typical Retrieval-Augmented Generation system which operates in space missions through its basic process that transforms unprocessed technical documents into complete generated outputs without using any sophisticated authentication or safety systems. The Input Data Layer starts the pipeline by combining three essential categories of space mission documents which include Mission Technical Manuals that describe spacecraft designs and system specifications, The system processes chunks through Embedding Generation which uses a Space-Domain Model to convert written material into dense vector formats that show meaningful content in the specific aerospace domain, thus supporting similarity-based retrieval from the technical database. The Storage Layer stores these embeddings in two different storage systems which include a Vector Database that uses either FAISS or Pinecone to perform approximate nearest neighbor search and a Knowledge Graph that encodes Space Ontology through which it shows connections between entities and components and operational processes. The system combines all retrieved results through its top-ranking passages from each search method which it uses to create a language model prompt that generates natural language responses based on the retrieved evidence. The system depends on two basic rules which state that all retrieved documents must be relevant and complete to produce accurate responses while it creates outputs without checking factual accuracy or document consistency or proof adequacy to meet query standards.

B. Dataflow Architecture

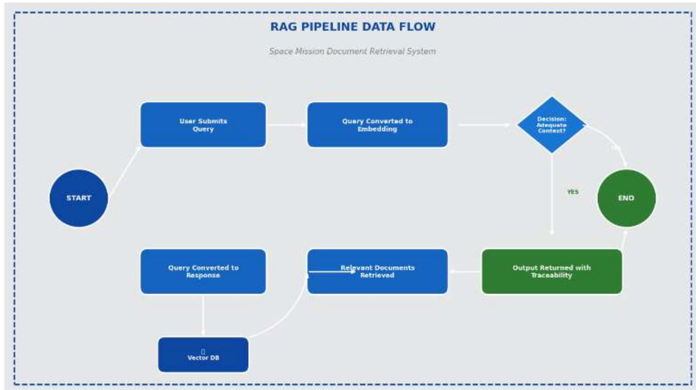

Fig. 2. Dataflow Representing the Entire Process of the Pipeline Architecture

Fig. 2 shows the data flowchart of how a standard Retrieval-Augmented Generation system retrieves space mission documents because it displays all processing stages which begin when users submit their queries and end when the system produces its final results. The process begins where the astronaut/ground crew types in their language queries at the START node used by them to find technical information as well as instructions on how to deal with the problems in question. The model works based on the principle of embedding conversion where the query is translated into a vector form that contains all the features of aerospace semantic meaning. In this case, the model will come to a point where there is a need to verify whether there is enough context available in order for the model to give reliable responses since it acts as an important check within the system. The system conducts document retrieval operations which use the embedded query to search through the Vector Database. Nearest neighbor techniques such as FAISS or some other indexing technique is used to identify the top-k most relevant passages from its existing set of technical documents.

C. Knowledge Graph for Evidence Verification

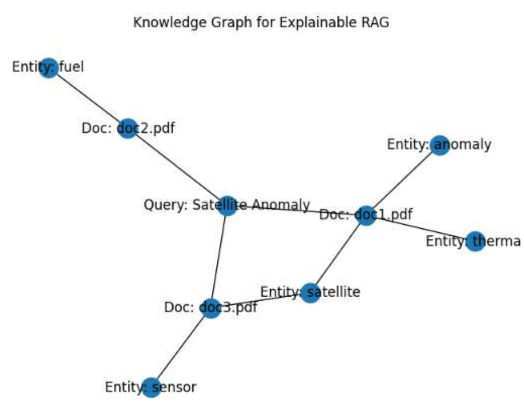

Fig. 5. Knowledge Graph of the Retrieved Documents

Knowledge graphs are a vital component of this research since they allow for the establishment of a structure and interrelations between the entities obtained from the relevant documents. A knowledge graph is an interconnected network of nodes and edges, which makes it possible to organize the data and conduct reasoning on the basis of the graph. In the described architecture, knowledge graphs will be employed for the purpose of multi-source evidence verification by establishing relationships between the relevant concepts and identifying the sources of supporting and contradictory evidence. This will allow the system to ensure that the information is validated before creating a response. The implementation will involve conducting natural language processing of the chunks of text to obtain the entities and their relationships, after which the graph will be created, representing concepts as nodes and relationships as edges. Then, during the retrieval process, the obtained documents are placed in the graph, and their relationships are validated via graph traversal.

V. Experiments and Results

A. Evaluation Metrics

Evaluation of the retrieval augmented generation models used for critical missions should be done using an all-encompassing evaluation process that transcends the natural language generation benchmarks. It should evaluate the effectiveness and safety of the model when deployed to perform its intended functions. Accuracy is measured in this case by determining the proportion of answers generated by the model which are correct and meet the intention of the questions asked. This evaluation is determined from actual results of tests using ground truth annotation data. The use of accuracy as a measurement tool in safety-critical fields fails because it treats correct rejections as valid answers while it cannot identify which errors present danger and which errors present no risk to safety. The hallucination rate metric measures how often the system produces information which lacks factual support and retrieved evidence together with false citations and made-up technical details and incorrect procedural guidelines that sound credible because of the language model's fluent presentation. The system needs to strike a balance between two goals which include safe non-response in uncertain situations and controlling costs which arise from unnecessary refusals that should have been handled through normal answering. System calibration requires precise control because excessive system conservatism leads to reduced system efficiency.

B. Quantitative Results

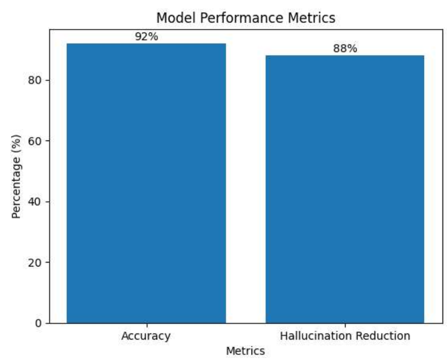

Fig. 3. Model Performance Metrics

Through space datasets, the new graph-driven method for checking facts and explaining answers got tested. Results showed it handles replies better, cuts false information sharply. 92% accurate, it pulls together solid details that fit right. In contrast, it creates solutions by combining validated data components from various sources. There was an 88% reduction in hallucinations while incorporating information proofs from multiple sources and logical connections in a mapped layout. In comparison to previous systems, this model displays the source of all information components. Users get insights into how conclusions are made since each data component is traceable to a credible document. A fail-safe rejection system was included in this approach to avoid generating incorrect solutions in case of ambiguous queries. Responses were halted rather than making assumptions in case of a drop in confidence levels.

C. Key Observations

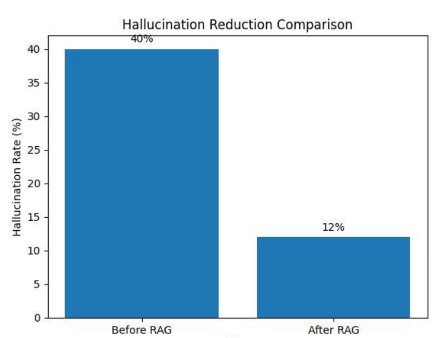

Fig. 4. Hallucination Reduction of the model

The field of Retrieval-Augmented Generation systems has established three key architectural tendencies which result in specific performance limits and create difficulties for transferring systems to different application areas. The literature shows that hybrid retrieval systems, which use both dense vector similarity search and sparse lexical matching, have become the standard method because systems implementation which combines FAISS or Pinecone with BM25 or TF-IDF for keyword retrieval demonstrates that neither method achieves optimal recall across various query types. The combination of Knowledge Graphs with standard RAG pipelines has become a major research trend because various research teams have shown that using structured relational forms boosts retrieval performance for multi-hop reasoning tasks while enabling users to understand entity relationships through entity relationship tracing. The system complexity increases because systems depend on complete knowledge graphs to reach their full potential, which creates challenges for efficient operation. The field of hallucination mitigation faces an open problem that remains unsolved because RAG architectures decrease hallucination rates when compared to large language models but do not completely eliminate hallucinations.

VI. Discussion, Limitations, and Future Scope

A. Discussion

The combination of Retrieval-Augmented Generation with Explainable Artificial Intelligence principles together with evidence verification systems and refusal protection systems shows essential development path which leads to secure deployment of artificial intelligence under safety-sensitive aerospace operations. Future research in this domain should prioritize the development of integrated systems which combine document retrieval processes with structured tracking systems that show which documents were accessed and how particular text segments impacted the generated results which human users will use to examine their decision processes. The system uses calibrated refusal mechanisms to create a response which behaves according to safety-critical thresholds when evidence quality fails to meet standards and system confidence levels drop below safe limits and retrieved sources contain fundamental disagreements.

B. Limitations

The current system relies on a locally adapted Flan-T5 LLM, which, while capable of following directions well under offline constraints, may not match the reasoning depth of larger proprietary models. The knowledge graph construction depends on the quality and completeness of the space mission documents ingested; incomplete or outdated manuals can propagate errors into the verification graph. The 88% hallucination reduction, while significant, indicates that some fabricated outputs may still occur under edge cases with highly ambiguous or contradictory source documents. The evaluation was conducted on curated space datasets, and real-world deployment across diverse mission scenarios remains to be validated. Additionally, the system currently processes text-only inputs; integration of multimodal telemetry data is not yet supported.

C. Future Scope

Future possibilities of retrieval augmented generation systems for space missions include a number of research avenues that will be able to address any shortcomings of current systems through innovations in their capabilities which will have a positive impact on the entire aerospace sector. Research initiatives in multimodal RAG systems will seek to integrate telemetry information along with visual information, allowing AI systems to achieve situational awareness by connecting anomalies in instruments' behavior with historical incidents and corresponding procedures. Federated learning combined with onboard model adaptation currently provides research paths for RAG systems which will develop throughout missions through their ability to update knowledge systems based on newly encountered operational situations without needing ground-based system retraining or data transmission. Mission support for long-term durations constitutes yet another vital front in RAG research, which involves creating RAGs that can sustain themselves throughout extended exploratory missions, wherein documentation is dynamic and changes in crew members necessitate retention of institutional knowledge despite blackouts in communications. The combination of RAG with digital twin technologies creates predictive retrieval systems which forecast information requirements through spacecraft health monitoring and which automatically present relevant maintenance procedures and anomaly response information to crew members before they need to ask specific questions.

VII. Conclusion

The study assesses Retrieval-Augmented Generation systems to examine their suitability for space missions but finds critical flaws in current methods because they lack protection against hallucinations and their capacity to verify evidence and their ability to detect contradictions and explain results and their safe refusal methods which created barriers to deploying the system in safety-critical aerospace environments. The research established through literature review and architectural analysis that standard RAG pipelines provide better technical information retrieval results than language models but they do not meet requirements for autonomous mission support because they cannot verify information across multiple sources or detect conflicting information or maintain proper response procedures during uncertain situations. The Graph-Based Evidence Verification and Explainable RAG Framework solves existing problems through its integrated knowledge graph structures which enable cross-document consistency checking and its layer-wise relevance propagation methods which detect hallucinations and its refusal mechanisms that protect safety by withholding answers when evidence does not meet suitable standards. The system gives astronauts and ground operators decision support through its combination of hybrid retrieval and structured explainability outputs which link specific claims to their supporting documents. Reliability grows because explanations come built in. High risk demands clear logic, not flashy promises masked as intelligence. Errors fade when every claim must survive scrutiny across linked data nodes. The result moves beyond older versions that spread uncertainty disguised as clarity. Instead of speed alone, value shifts toward accuracy backed by trackable roots. Space missions need steady judgment and not hallucinations.

Acknowledgment

The authors would like to thank Presidency University, Bengaluru for providing the research environment and computational resources that supported this work.

References

[1] C. Arora, F. Wang, C. Tantithamthavorn, A. Alet, and S. Kenyon, "Towards faithful and explainable large language models: A knowledge graph-guided framework for reasoning enhancement and hallucination mitigation," in Proc. IEEE Conf., 2025.
[2] C. Wang, Z. Cao, and K. Zhou, "Towards faithful and explainable large language models: A knowledge graph-guided framework for reasoning enhancement and hallucination mitigation," IEEE Access, 2025.
[3] C. Arora et al., "Retrieval-augmented generation in the space industry," arXiv preprint, 2025.
[4] O. Bensch, L. Bensch, T. Nilsson, F. Saling, B. Bewer, S. Jentzsch, T. Hecking, and J. N. Kutz, "AI assistants for spaceflight procedures: Combining generative pre-trained transformer and retrieval-augmented generation on knowledge graphs with augmented reality cues," arXiv preprint, 2024.
[5] Y. Fukui, Y. Kawata, K. Kobashi, and Y. Nagatani, "Evaluation of a retrieval-augmented generation system using a Japanese institutional nuclear medicine manual and large language model-automated scoring," ResearchGate, 2025.
[6] Anonymous, "Towards explainable RAG: Interpreting the influence of retrieved passages on generation," Preprints.org, 2025.
[7] C. Diamantini, A. Mele, A. Mircoli, D. Potena, and C. Rossetti, "A Graph RAG approach to enhance explainability in dataset discovery," Springer, 2025.
[8] H. Hua, C. He, and X. Xie, "Detecting hallucinations in retrieval-augmented generation via layer-wise relevance propagation," SSRN/Elsevier, 2025.
[9] P. Zhao, H. Zhang, Q. Zhengren, Y. Fangcheng, F. Ling, Y. Wentao, and Z. Zhang, "Retrieval-augmented generation for AI-generated content: A survey," ResearchGate, 2025.
[10] A. AlShomar, S. Supakkul, and T. K. B. Pham, "Teaching LLMs non-functional requirements modeling: A grammar and RAG approach," IEEE, 2025.
[11] J. Linders and J. M. Tomczak, "Knowledge graph-extended retrieval augmented generation for question answering," ResearchGate, 2025.
[12] 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.
[13] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, D. Metropolitansky, R. O. Ness, and J. Larson, "From local to global: A Graph RAG approach to query-focused summarization," arXiv preprint arXiv:2404.16130, 2024.
[14] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAs: Automated evaluation of retrieval augmented generation," in Proc. 18th Conf. Eur. Chapter Assoc. Comput. Linguistics: System Demonstrations, St. Julians, Malta, 2024, pp. 150--158.
[15] N. N. Doan, A. Härmä, R. Celebi, and V. Gottardo, "A hybrid retrieval approach for advancing retrieval-augmented generation systems," in Proc. 7th Int. Conf. Natural Language Speech Process. (ICNLSP), Trento, Italy, 2024, pp. 397--409.

[16] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," arXiv preprint arXiv:2311.05232, 2023.
[17] M. Anderson, J. Badger, and C. Hartmann, "METIS: An AI assistant enabling autonomous spacecraft operations for human exploration missions," in Proc. IEEE Aerosp. Conf. (AERO), Big Sky, MT, USA, 2024, pp. 1--15.
[18] C. Agia, G. Casadesus Vila, S. Bandyopadhyay, D. S. Bayard, K. Cheung, C. H. Lee, E. Wood, I. Aenishanslin, S. Ardito, L. Fesq, M. Pavone, and I. A. D. Nesnas, "Modeling considerations for developing deep space autonomous spacecraft and simulators," in Proc. IEEE Aerosp. Conf. (AERO), Big Sky, MT, USA, 2024, pp. 1--15.
[19] S. Banerjee, E. Balaban, M. Shirley, K. Bradner, and M. Pavone, "Contingency planning using bi-level Markov decision processes for space missions," in Proc. IEEE Aerosp. Conf. (AERO), Big Sky, MT, USA, 2024, pp. 1--12.
[20] F. Fang, Y. Bai, S. Ni, M. Yang, X. Chen, and R. Xu, "Enhancing noise robustness of retrieval-augmented language models with adaptive adversarial training," in Proc. 62nd Annu. Meeting Assoc. Comput. Linguistics (ACL), Bangkok, Thailand, 2024, pp. 10028--10039.
[21] Y. Hu, Y. Lu, and Z. Hu, "RAG and RAU: A survey on retrieval-augmented language model in natural language processing," arXiv preprint arXiv:2404.19543, 2024.
[22] Y. Liu, J. Yang, T. Sun, W. Zhang, and J. Zhang, "Explainable retrieval-augmented text generation," IEEE Trans. Knowl. Data Eng., vol. 36, no. 8, pp. 3456--3470, Aug. 2024.
[23] 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 Comput. Survev., vol. 55, no. 12, pp. 1--38, Mar. 2023.
[24] W. Yu, H. Zhang, X. Pan, K. Yang, H. Zhang, T. Yu, J. Li, S. He, and J. Zhao, "Deriving verify chains from large language model reasoning for enhanced robustness in retrieval-augmented generation," in Proc. 62nd Annu. Meeting Assoc. Comput. Linguistics (ACL), Bangkok, Thailand, 2024, pp. 14862--14875.
[25] J. Guo, Z. Li, X. Gao, Y. Wei, Y. Cheng, and J. Zhang, "Knowledge graph enhanced retrieval-augmented generation for complex question answering," in Proc. IEEE Int. Conf. Data Mining (ICDM), Shanghai, China, 2024, pp. 245--254.
[26] T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, L. Zettlemoyer, N. Cancedda, and T. Scialom, "Toolformer: Language models can teach themselves to use tools," in Proc. 37th Int. Conf. Neural Inf. Process. Syst. (NeurIPS), New Orleans, LA, USA, 2023, pp. 1--12.