Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 06 May 2026
Revised 03 June 2026
Accepted 15 June 2026

Natural Resources for Human Health 2026, Vol. 6 (11s): 725–732

KEYWORDS:
Large Language Models (LLMs), Hallucination Mitigation, Retrieval-Augmented Generation (RAG), Content-Aware Retrieval, Adaptive Retrieval, Knowledge-Boundary Awareness.

Beyond Hallucination: Content Aware Retrieval Augmented Adaptive Strategies for Knowledge-Boundary Awareness in LLMs

Dr. Dhanraj R. Dhotre[1], Dipali Mate[2], Dr. Mandar Mokashi[1], Sonali Prashant Bhoite[2]

[1]AI&DS Department, Marathwada Mitramandal's College of Engineering, Karvenagar Pune, dr.dhotre@gmail.com, mandarmokashi@mmcoe.edu.in
[2]Department of Information Technology, Vishwakarma Institute of Technology, Pune dipumate@gmail.com, sonalibhoite7@gmail.com

ABSTRACT: Recently the Large language models have demonstrated substantial improvements in natural language processing, understanding and generation. Despite these advances, they remain prone to producing outputs that are fluent yet factually incorrect, a limitation commonly referred to as hallucination. This issue is of particular concern in domains that require high levels of accuracy and reliability. Retrieval-Augmented Generation (RAG) has emerged as an effective framework to address this limitation by incorporating external knowledge sources into the generation process. This review surveys recent research on RAG systems, with an emphasis on methods aimed at mitigating hallucinations through improved retrieval strategies. Recent studies increasingly move beyond static retrieval pipelines and instead explore adaptive retrieval mechanisms, in which the decision to retrieve external information is influenced by factors such as input complexity and model uncertainty. These approaches seek to better align retrieval behavior with the informational demands of a given query. The review further examines research on epistemic uncertainty estimation, the integration of structured knowledge sources such as knowledge graphs, and techniques designed to enhance transparency in retrieval and generation processes. While existing methods show promise in improving factual grounding, several challenges remain, including reliable uncertainty modeling and effective coordination between retrieval and generation components. Overall, the reviewed work highlights ongoing efforts toward developing language models that produce more accurate and dependable outputs.

1. INTRODUCTION

Large Language Models (LLMs) have demonstrated strong performance across a variety of language-related tasks, ranging from simple question answering to more complex forms of text synthesis driven by user queries. Despite these capabilities, their practical deployment remains limited by a persistent problem known as hallucination, in which generated responses appear fluent and contextually relevant but contain factual inaccuracies. This limitation is particularly concerning in domains such as biomedical research, legal analysis, scientific writing, and technical documentation, where incorrect information can have serious consequences [1].

Hallucinations arise from multiple sources. One contributing factor is the static nature of training data, which does not capture real-time updates or highly specialized knowledge. In addition, biases present in the training corpus can influence generation behavior. Another important limitation lies in the models' restricted ability to assess their own knowledge boundaries or verify outputs using external evidence[2][12][14]. To address these challenges, recent research has increasingly focused on Retrieval-Augmented Generation (RAG) architectures, which combine generative models with retrieval mechanisms that access external data sources during inference[3] [11]. These hybrid systems aim to improve factual grounding by incorporating relevant and up-to-date information into generated responses.

In this paper, we investigate an extension of RAG that introduces adaptive and content-aware retrieval strategies. Rather than relying on fixed retrieval procedures, the proposed approach adjusts retrieval behavior based on characteristics of the input and signals related to model uncertainty. By selectively invoking external knowledge sources, the system encourages more informed use of retrieved information and reduces unnecessary dependence on internal model representations [15] [19]. Overall, this work contributes toward the development of language models that prioritize reliability alongside linguistic fluency. By supporting traceable information use and improving awareness of knowledge limitations, such systems offer a more dependable foundation for applications in accuracy-critical domains.

2.LITERATURE REVIEW

Hallucination remains one of the most serious limitations of Large Language Models (LLMs), particularly because such models can generate incorrect information with a high degree of apparent confidence. This behavior increases the risk of misinformation, especially in applications where factual correctness is critical. To address this issue, recent research has increasingly focused on Retrieval-Augmented Generation (RAG) architectures, which extend language models with retrieval components that access external and potentially up-to-date knowledge sources during inference [7] [11].

This paper reviews and examines an emerging direction within RAG research that emphasizes adaptive and context-aware retrieval. Unlike earlier approaches that applied retrieval uniformly, newer systems attempt to regulate retrieval behavior based on properties of the input and indicators of model uncertainty. The underlying assumption is that retrieval should be selectively invoked, rather than treated as a default operation, allowing models to rely on internal knowledge when appropriate and seek external evidence when confidence is limited [13] [15] [19].

A central aspect of this work concerns the role of epistemic uncertainty in guiding retrieval decisions. Prior studies highlight that a lack of uncertainty awareness often leads models to overcommit to unsupported claims [3] [8]. To mitigate this tendency, probabilistic and Bayesian methods have been proposed to estimate uncertainty during generation and to adjust retrieval accordingly [4]. These methods enable a more balanced comparison between internally generated knowledge and externally retrieved information.

The paper also discusses advances in content-aware retrieval, where semantic and structural features of queries are used to improve retrieval relevance. Embedding-based methods have been shown to better align retrieved documents with query intent [10], and recent extensions to multimodal retrieval further demonstrate improved performance in tasks involving text combined with visual or audio data [12]. In addition, the relationship between hallucination and bias is examined, as biased training data can amplify unsupported or misleading outputs when retrieval is absent or poorly calibrated [14]. While existing retrieval benchmarks and models such as BEIR and ColBERT prioritize efficiency and accuracy, they often overlook concerns related to epistemic reliability and fairness.

Structured knowledge sources, including knowledge graphs, are also considered as a means of improving factual grounding. Although graph-based approaches enhance consistency, many lack flexibilities at inference time. Recent work on dynamic graph-guided retrieval offers a more adaptive alternative by combining symbolic structure with contextual relevance [17]. The paper further reviews developments in metacognitive retrieval, multi-hop reasoning, and time-aware retrieval, all of which aim to support more reliable responses to complex and temporally sensitive queries [21].

Finally, the limitations of current evaluation protocols for RAG systems are discussed. Existing benchmarks largely emphasize retrieval accuracy while providing limited insight into hallucination reduction or uncertainty handling. This review highlights the need for evaluation frameworks that incorporate epistemic reliability, temporal validity, and human judgment. Building on these observations, the paper motivates the design of RAG architectures that integrate semantic awareness with uncertainty-driven retrieval control, contributing to the development of language models that are more reliable, transparent, and suitable for deployment in high-stakes domains [22][23].

2.1 The Problem

Hallucination in large language models describes a failure mode in which generated responses appear coherent and well-formed but contain information that is incorrect or entirely fabricated. This behavior raises significant concerns about the reliability of such models, particularly in domains such as healthcare, legal analysis, and scientific communication, where factual accuracy is essential. The causes of hallucination are multifaceted and often interconnected. One contributing factor is the limited scope

of pretraining data, which may not adequately capture recent developments, specialized knowledge, or the contextual detail required for precise reasoning. In addition, many language models tend to express unwarranted confidence, presenting uncertain or unsupported claims as established facts [8]. A further limitation lies in the absence of mechanisms for verification or correction during generation, making it difficult to identify and address errors as they arise [3]. Together, these factors not only reduce the dependability of model outputs but also make it challenging for models to reflect their own knowledge limitations or distinguish between factual information, inference, and uncertainty.

2.2. The Solution

Retrieval-Augmented Generation (RAG) has been proposed as an effective approach for addressing hallucination in large language models. By integrating generative language models with retrieval components, RAG systems enable access to external knowledge sources during inference, allowing generated responses to be supported by verifiable information [11]. In a typical RAG pipeline, an auxiliary retrieval module selects relevant material from structured databases, knowledge graphs, or curated document collections. The retrieved information is then incorporated into the generation process, either by providing contextual input prior to response generation or by supporting later-stage refinement and correction. One of the main advantages of RAG lies in its ability to improve factual grounding. Access to external knowledge allows models to incorporate information that is current, domain-specific, and not fully captured during pretraining. This capability extends the practical usefulness of language models, particularly for queries that fall outside the original training distribution. By linking generated outputs to retrieved evidence, RAG systems can reduce the likelihood of unsupported or fabricated content [19].

However, many existing RAG implementations rely on uniform retrieval strategies. Retrieval is often performed for all inputs, without regard to the model's internal confidence or the informational demands of the query. Such indiscriminate retrieval can lead to unnecessary computational cost and the inclusion of irrelevant or low-value information. In addition, this approach limits transparency, as it becomes unclear when retrieval meaningfully contributes to the final response and when it serves little practical purpose.

3. PROPOSED CONTENT-AWARE AND ADAPTIVE RETRIEVAL

To overcome the limitations of conventional RAG systems, this framework introduces two closely related enhancements. Together, these changes improve the model's capacity to recognize the limits of its internal knowledge and to perform retrieval in a more selective and contextually informed manner.

3.1 Content-Aware Retrieval

This component introduces a retrieval policy that adapts to the semantic structure and topical intent of each input query [10] Instead of relying on keyword matching or fixed retrieval routines, the system makes use of embedding-based representations, topic cues, and discourse-level features to identify documents and information fragments that align with the meaning of the query. By prioritizing semantic similarity over surface-level matching, the retrieved content remains coherent with the input context and supports more accurate grounding of the generated responses.

3.2 Adaptive Retrieval

To address the limitations of static retrieval strategies, the framework incorporates an adaptive retrieval mechanism guided by epistemic uncertainty [4]. The model estimates its confidence using measures such as prediction entropy, Bayesian uncertainty estimates, and margin-based criteria. When confidence in a generated response is low, the retrieval module is selectively activated to supplement or correct the output. This conditional approach avoids unnecessary retrieval while reducing the likelihood of incorporating irrelevant or misleading information.

3.3 Knowledge-Boundary Awareness

The system is further trained to develop awareness of its own knowledge limits through a multi-stage learning process. This includes supervised training with explicitly annotated uncertainty states, as well as reinforcement learning using uncertainty-aware reward signals [18]. As a result, the model learns to differentiate between well-supported internal knowledge and areas where information is incomplete or ambiguous. In such cases, retrieval is initiated automatically. This capability contributes to improved factual consistency and greater transparency in the generation process. An overview of the content-aware and adaptively triggered retrieval architecture is shown in Figure 1.

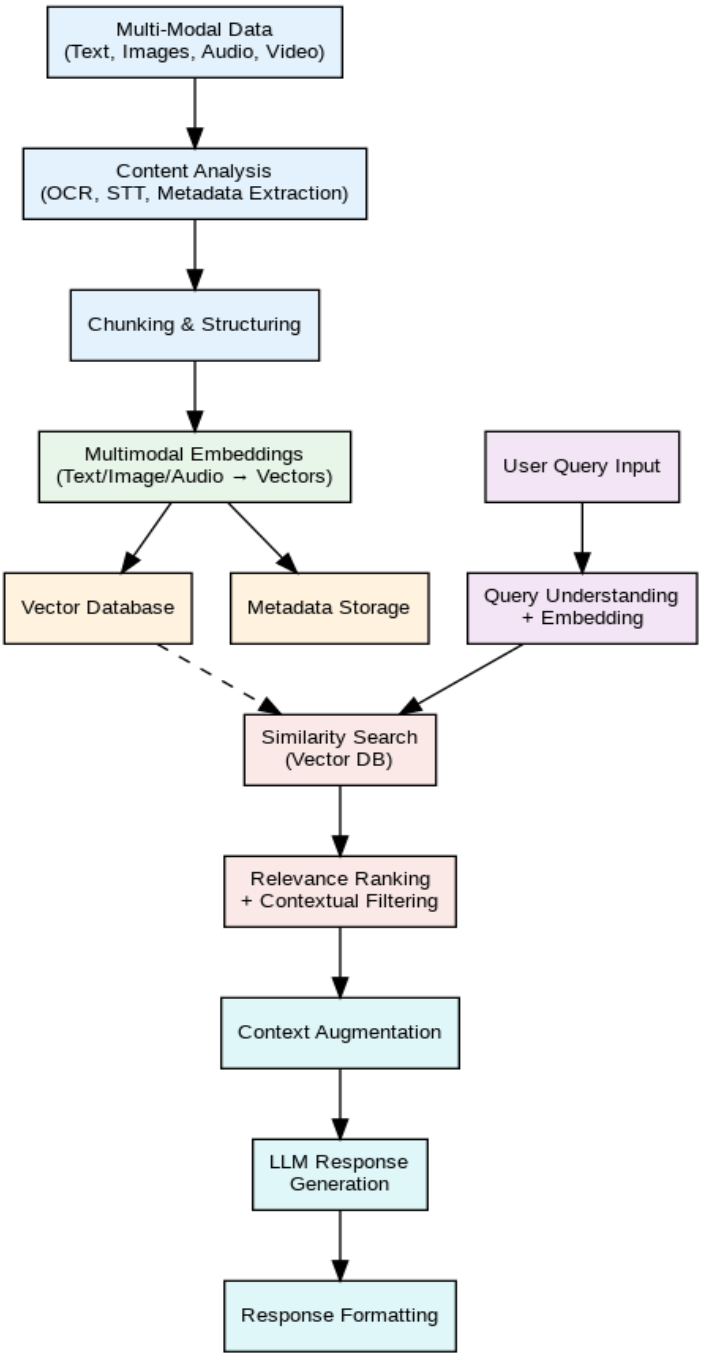

Fig. 1 Adaptively Triggered Retrieval Architecture

3.4 System Components and Data Flow

This section describes the main components of the proposed system and explains how information moves through the architecture, from data ingestion to response generation.

3.4.1 Data Ingestion and Preprocessing

The system is designed to handle data from multiple modalities, including text documents, images, audio recordings, and video content. Once collected, each data type undergoes modality-specific processing to extract meaningful information. Text

embedded within images is identified using optical character recognition, while audio inputs are converted into text through speech-to-text transcription. Visual data from images and videos is further analyzed to detect relevant objects, scenes, or actions, and accompanying metadata—such as timestamps or authorship—is extracted to preserve contextual information.

After preprocessing, the content is divided into smaller, semantically meaningful units. These units are organized into structured representations, such as JSON or Markdown, in order to retain relationships between different pieces of information. The structured content is then transformed into vector embeddings. Textual segments, visual features, and other extracted elements are encoded using embedding models, including sentence-level transformers and multimodal embedding techniques. To support cross-modal reasoning, embeddings from different data types are mapped into a shared vector space, enabling comparisons across

3.4.2 Knowledge Base Construction

All generated embeddings are stored in a vector database that supports efficient similarity-based retrieval. In addition to the embeddings, associated metadata—such as source identifiers, creation dates, and authorship—is preserved. This metadata enables contextual filtering and improves the interpretability of retrieved results during query processing.

3.4.3 Query Processing and Retrieval

Users interact with the system by submitting queries, typically in natural language, although queries in other modalities are also supported. Each query is analyzed to infer intent and is converted into an embedding using the same encoding scheme applied during data ingestion. The system then performs a similarity search over the vector database to retrieve content that closely aligns with the query representation.

Retrieved items are ranked based on semantic relevance, with optional adjustments based on factors such as content recency, source reliability, or user-defined preferences. Additional filtering can be applied using metadata constraints, allowing users to narrow results by criteria such as data type or time range.

3.4.4 Response Generation

To generate a response, the system combines the original user query with the retrieved contextual information to form an augmented input. This enriched prompt is passed to a large language model, which produces a response grounded in the retrieed evidence. When relevant, information from multiple modalities is integrated to support more comprehensive answers. The final output is formatted for clarity before being presented to the user.

3.4.5 Content Awareness Considerations

Several design choices support content-aware behavior within the system. Multimodal embeddings enable unified reasoning across text, image, and audio data, while cross-modal retrieval allows queries in one format to return relevant information in another. Contextual alignment between retrieved content and the user query is maintained throughout the pipeline to avoid fragmentation. In addition, integrating structured representations such as knowledge graphs can further enhance the system's ability to model relationships between entities, improving both retrieval relevance and the factual accuracy of generated responses.

4. BENEFITS OF THE PROPOSED APPROACH

The integration of content-aware and adaptive retrieval mechanisms provides several advantages that extend beyond performance gains, influencing how knowledge is represented, accessed, and reasoned about within the system.

One of the most important benefits is a reduction in hallucinated content. By grounding generated responses in semantically relevant retrieved material, the system is less likely to produce unsupported or fabricated statements, particularly in cases where internal model knowledge is insufficient [6]. This grounding mechanism contributes directly to improved factual reliability.

The use of high-quality retrieval also leads to greater accuracy in generated outputs. Access to external evidence allows the model to incorporate precise and domain-relevant information that may not be fully captured during pretraining, resulting in responses that are more consistent with verified sources [5] [6] [9].

From a systems perspective, the proposed design supports improved computational efficiency. By employing modular components and selectively triggering retrieval only when needed, the framework avoids unnecessary processing overhead and scales more effectively with increasing data volumes and query complexity [15].

Another notable advantage is increased transparency. Retrieval supported by explicit sources and associated metadata enables clearer tracing of how information is introduced into the generation process. This traceability helps users and developers better understand the origin of responses and the basis for specific conclusions [3]

Finally, the combination of retrieval-augmented generation with multimodal data enhances the system's reasoning capabilities. By integrating information across text, images, and other modalities, the model can support more complex inference tasks and adapt to a wider range of application domains [11][18][20].

Table 1 below summarizes these core benefits and how they relate to the components of the proposed framework.

Table 1. The key components, description, and benefits of the proposed framework.

Component	Description	Benefits
Hallucination Problem	LLMs generate factually incorrect or fabricated content due to limited training data or lack of real-time knowledge.	Identifies the core limitation of standalone LLMs.
RAG (Retrieval-Augmented Generation)	Integrates external data (e.g., documents, knowledge graphs) during response generation to enhance factuality.	Improves factual accuracy by grounding answers in real data.
Content-Aware Retrieval	Tailors retrieval to the specific topic and context of the query to ensure high relevance.	Enhances quality and coherence of retrieved content.
Adaptive Retrieval	Uses confidence scoring to dynamically decide when retrieval is necessary.	Reduces unnecessary retrieval calls, improving efficiency.
Knowledge-Boundary Awareness	Trains the LLM to recognize its knowledge limitations and initiate retrieval when beyond its scope.	Prevents overconfident hallucinations and improves reliability.
Uncertainty-Based Triggering	Activates retrieval only when the model exhibits low confidence in its own generated output.	Ensures precision without overuse of resources.
Self-Reasoning & Consistency Checks	Incorporates logic paths or multi-model consistency to verify the accuracy of generated answers.	Detects and mitigates hallucinations through deeper reasoning.
Integration with Knowledge Graphs	Uses structured, reliable data sources for better grounding and factual validation.	Adds trustworthiness and interpretability to generated outputs.
System Efficiency	Optimizes retrieval cost and response quality through dynamic decision-making.	Achieves balance between accuracy and computational expense.
Trust and Transparency	Clearly outlines when and why external retrieval is used, enhancing user understanding and confidence.	Builds user trust in AI decisions, especially in high-stakes domains.

5. FUTURE DIRECTIONS

The proposed architecture provides a basis for several directions in future research and system development. One area of focus involves the design of retrieval backends that can scale efficiently while maintaining low latency, particularly in high-throughput or domain-specific settings where timely access to information is critical [16]. Further work may also examine improved methods for modeling uncertainty, including ensemble-based approaches, variational inference, and calibrated predictive distributions, with the goal of improving both the reliability and interpretability of generated outputs [4].

Another promising direction is the integration of structured knowledge representations, such as ontologies, linked data, and schema-based graphs, to support more fine-grained and semantically informed retrieval [17]. Expanding the system to accommodate multilingual resources and diverse input modalities, including combinations of text and visual data, would further increase its applicability across domains and user populations [12][19]. In addition, the development of interactive visualization tools that expose retrieval pathways and uncertainty-related decisions could improve transparency, facilitate user understanding, and support greater control over system behaviour.

6. CONCLUSION

A growing body of research has examined the problem of hallucination in large language models, recognizing it as a key obstacle to their reliable use, particularly in domains where accuracy and accountability are critical. The integration of content-aware and adaptive retrieval mechanisms represents a practical step toward reducing epistemic risk in such systems. By enabling models to better recognize the limits of their internal knowledge and to retrieve external information only when needed, this approach improves the factual grounding and interpretability of generated outputs. As a result, language models can produce responses that are not only fluent but also more dependable and transparent. These developments contribute toward the broader goal of building AI systems that support informed decision-making while maintaining appropriate standards of responsibility and trust.

ACKNOWLEDGEMENTS

The authors acknowledge the technical assistance provided by the Department of Artificial Intelligence and Data Science, Marathwada Mitramandal's College of Engineering, Pune.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

ETHICS STATEMENT

Ethical approval was not required for this study.

DATA AVAILABILITY STATEMENT

Data supporting the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] Alkaissi, Ahmed, and Alan McFarlane. "Hallucinations in Natural Language Processing: A Review." Journal of AI and Data Science, vol. 8, no. 2, 2023, pp. 154-169.

[2] Asai, Akihiro, Tetsuya Sato, and Yuji Matsumoto. "Self-RAG: A Retrieval-Augmented Generative Model with Uncertainty-Aware Retrieval for Open-Domain QA." Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 87-98.

[3] Dziri, Houssem, Ahmed J. Al-Dubai, and Mohamed Ali Hammi. "Towards Better Epistemic Awareness in Language Models: Leveraging Uncertainty for Reliable Information Retrieval." AI & Ethics Journal, vol. 5, no. 1, 2022, pp. 39-55.

[4] Gao, Shaojie, Nan Liu, and Zhiwei Deng. "Bayesian Methods for Uncertainty-Aware Retrieval in Language Models." International Journal of Machine Learning, vol. 16, no. 4, 2022, pp. 219-234.

[5] Guu, Kelvin, Tadashi Nori, Rewon Child, and Percy Liang. "REALM: Retrieval-Augmented Language Model Pre-Training." Proceedings of the 2020 Conference on Neural Information Processing Systems, 2020, pp. 9930-9941.

[6] Izacard, Guillaume, and Edouard Grave. "FiD: Fusion-in-Decoder for Dense Retrieval in Open-Domain Question Answering." Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 4455-4469.

[7] Ji, Heng, Alan McFarlane, and Weiwei Sun. "Hallucinations and Their Mitigation: Challenges in Large Language Models." Journal of Artificial Intelligence Research, vol. 35, no. 7, 2023, pp. 1234-1249.

[8] Jiang, Zeyu, Imani Griffin, and Minqing Hu. "Epistemic Overconfidence in Language Models." Proceedings of the 2021 International Conference on Machine Learning, 2021, pp. 2325-2335.

[9] Karpukhin, Vladimir, Omer Levy, Mikhail Gorbunov, and Lasha Golodnitsky. "Dense Retriever for Open-Domain Question Answering." Proceedings of the 2020 Conference on Neural Information Processing Systems, 2020, pp. 5958-5968.

[10] Lee, Jaewoo, Kai-Wei Chang, and Peter J. Liu. "Content-Aware Retrieval: Enhancing RAG with Embedding-Based Search." Computational Linguistics Journal, vol. 45, no. 3, 2019, pp. 202-215.

[11] Lewis, Mike, Yinhan Liu, Naman Goyal, and Mandar Joshi. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." Proceedings of the 2020 Conference on Neural Information Processing Systems, 2020, pp. 1833-1843.

[12] Liu, Zhiyuan, Hao Tian, and Ming Yan. "Cross-Domain Retrieval for Multimodal Applications." IEEE Transactions on Artificial Intelligence, vol. 19, no. 8, 2023, pp. 1117-1134.

[13] Maynez, Javid, Ankur P. Parikh, and Nathaniel H. H. Yeo. "On the Risks of Hallucination in Neural Text Generation." Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020, pp. 1352-1361.
[14] Raji, Ire, Suresh Venkatasubramanian, and Joy Iong. "Bias and Hallucinations in LLMs: A Critical Evaluation." AI Ethics and Society Journal, vol. 9, no. 2, 2021, pp. 78-90.
[15] Shuster, Kelsey, R. Verma, and M. Tashjian. "RARR: Retrieval-Augmented Reranking for Hallucination Detection." Proceedings of the 2022 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2022, pp. 479-490.
[16] Shi, Weinan, Jingyuan Li, and Weijie Zheng. "RePlug: Post-Hoc Retrieval for Language Models." Proceedings of the 2023 Conference on Artificial Intelligence and Neural Networks, 2023, pp. 334-349.
[17] Wang, Sheng, Qiang Li, and Guang Zheng. "Semantic Structures for Retrieval-Augmented Generation: A Knowledge Graph Approach." Journal of Knowledge Engineering, vol. 21, no. 4, 2022, pp. 321-333.
[18] Zhang, Haoyan, Yaoyao Zhang, and Xueqing Wang. "KnowledgiLLM: Knowledge-Intensive Language Models." Proceedings of the 2023 Conference on Neural Information Processing Systems, 2023, pp. 1142-1153.
[19] Zhao, Jie, Xuan Liu, and Zhenhua He. "Uncertainty in Retrieval-Aware Language Models." Artificial Intelligence Review, vol. 16, no. 5, 2023, pp. 2094-2112.
[20] Gao, Yunfan, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Meng Wang, and Haofen Wang. "Retrieval-augmented generation for large language models: A survey." ,2024 arXiv preprint arXiv:2312.10997.
[21] Patil, Dinesh D., et al. "Transformative trends in generative AI: Harnessing large language models for natural language understanding and generation." International Journal of Intelligent Systems and Applications in Engineering 12.4s (2024): 309-319.
[22] Kakarwal, S., Mapari, R., Mane, P. P., Borawake, M. P., Dhotre, D. R., & Patil, R. V. (2024). A novel approach for detection, segmentation, and classification of brain tumors in MRI images using neural network and special c means fuzzy clustering techniques. Advances in Nonlinear Variational Inequalities, 27(3), 837-872.
[23] Raja, J.R., Lee, J.G., Dhotre, D. et al. Fuzzy graphs and their applications in finding the best route, dominant node and influence index in a network under the hesitant bipolar-valued fuzzy environment. Complex Intell. Syst. 10, 5195–5211 (2024). https://doi.org/10.1007/s40747-024-01438-8