Joint Modeling of Intelligent Retrieval-Augmented Generation in LLM-Based Knowledge Fusion

Di Wu and Shuaidong Pan *

Posted Date: 10 September 2025

doi: 10.20944/preprints202509.0871.v1

Keywords: retrieval-enhanced generation; intelligent search algorithm; semantic alignment; knowledge fusion

Article

Joint Modeling of Intelligent Retrieval-Augmented Generation in LLM-Based Knowledge Fusion

Di Wu [1] and Shuaidong Pan [2,*]

[1] University of Southern California, Los Angeles, USA

[2] Carnegie Mellon University, Pittsburgh, USA

* Correspondence: shuaidongpan@gmail.com

Abstract

This study addresses the insufficient connection between retrieval and generation in large-scale knowledge utilization by proposing a retrieval-augmented generation method enhanced with intelligent search algorithms. The approach encodes input queries and candidate knowledge passages into a unified semantic space and dynamically aggregates relevant knowledge through similarity measures and attention weighting, ensuring that the generation stage receives high-quality external knowledge support. A fusion module is then constructed to jointly model retrieval and query representations, enabling the generation model to dynamically use retrieved content during text generation and ensuring semantic coverage, factual consistency, and contextual coherence. A joint optimization mechanism is further introduced to simultaneously optimize retrieval loss and generation loss, strengthening the interaction between the two modules and improving overall system performance. To validate the framework, comparative experiments were conducted on a publicly available discriminative dataset, along with sensitivity analyses under different hyperparameter settings, data perturbations, and environmental configurations. The experimental results show that the proposed method outperforms baseline models on key metrics such as F1, Precision, Recall, and ACC, while also demonstrating stability and robustness across vector dimensionality, similarity measures, and retrieval index scales. These findings confirm that the proposed framework can provide more accurate, comprehensive, and consistent knowledge support in complex contexts, establishing a solid foundation for advancing integrated research on retrieval and generation.

Keywords: retrieval-enhanced generation; intelligent search algorithm; semantic alignment; knowledge fusion

I. Introduction

In today's era of information explosion, the way knowledge is acquired and used is undergoing profound change. With the rapid growth of data and the increasing diversity of information types, knowledge from a single source or a single modality is no longer sufficient for complex tasks. Traditional information retrieval systems have played an important role in locating information quickly [1]. However, they often struggle to fully capture context, to identify cross-modal semantic relations, and to integrate fragmented knowledge resources effectively. The rise of generative models in natural language processing has brought knowledge utilization into a new stage. These models perform well in natural language generation and complex semantic expression. Yet they still face limitations in factual accuracy, external knowledge usage, and cross-domain adaptation [2–4]. Establishing a close connection between retrieval and generation, and improving retrieval-augmented generation through intelligent search algorithms, has become a key direction in artificial intelligence research and application[5].

 doi:10.20944/preprints202509.0871.v1

The core of retrieval-augmented generation lies in using a retrieval module to introduce dynamic external knowledge into the generation model[6]. This compensates for the static nature of model parameters and the lag in knowledge updating. The mechanism ensures that generated content is more accurate and informative at the factual level. It also provides advantages in knowledge coverage, contextual consistency, and cross-domain extension. Current methods, however, still need improvement in refined retrieval strategies [7], multimodal information fusion[8,9], and deep coupling between retrieval and generation. When facing multi-source heterogeneous data, traditional retrieval often depends on keywords or vector similarity and ignores semantic alignment between different modalities. Without high-quality knowledge input, generative models are prone to redundancy, hallucinations, or inconsistency. Introducing intelligent search algorithms and building an optimized framework for retrieval-augmented generation can overcome these bottlenecks. It can also drive the co-evolution of retrieval and generation at a higher level[10].

The development of intelligent search algorithms brings new opportunities for retrieval-augmented generation[11]. From heuristic search to learning-based search strategies, and from structured queries to unstructured semantic matching, the evolution of search algorithms reflects continuous improvements in information organization, relevance measurement, and contextual adaptation [12]. Search algorithms must balance efficiency and accuracy while dealing with large knowledge bases, varied data quality, and diverse user needs. When these capabilities are deeply integrated with generative models, the selection and use of knowledge can be more intelligent. The generated content becomes more logical, targeted, and controllable. Such integration is of theoretical value in research and has wide applications in practice. Examples include intelligent question answering, decision support, medical report generation, financial information analysis, and multimodal interaction systems[13].

Studying retrieval-augmented generation enhanced by intelligent search algorithms carries important scientific and practical significance. From an academic perspective, this research promotes deeper integration of information retrieval and natural language generation. It explores mechanisms for cross-modal and cross-domain knowledge representation and reasoning[14,15]. It extends the boundary of artificial intelligence in complex contextual understanding and generation. From an application perspective, as human-computer interaction grows more complex, users no longer accept static retrieval results or one-way text generation. They expect systems to dynamically combine information discovery and expressive generation. This integration improves factuality and personalization of content. It also reduces user effort in filtering and verifying knowledge, bringing new productivity tools and innovation models to knowledge-intensive industries.

In conclusion, the integration of retrieval-augmented generation and intelligent search algorithms represents an important trend in artificial intelligence research. It addresses the urgent need for large-scale knowledge utilization and aligns with the technological evolution of the intelligent information era. A framework with precision, adaptability, and interpretability can drive the joint optimization of knowledge retrieval and generation. It also lays a foundation for more intelligent and trustworthy human-computer interaction. This research not only broadens the theoretical scope of artificial intelligence but also creates long-term impact in information processing, knowledge services, and intelligent decision-making.

II. Method

At the methodological level, this study employs the selective knowledge injection technique introduced by Zheng et al. [16], which enables precise and efficient knowledge augmentation within large-scale language models. By applying Zheng et al.'s adapter-based approach, the framework constructs a unified semantic space for both queries and candidate passages, allowing for dynamic and context-aware knowledge selection during the retrieval process.

In designing the model's fusion and encoding mechanisms, we draw upon the time-aware and multi-source feature fusion methods proposed by Wang [17]. These techniques are incorporated to enhance the adaptability of semantic representations and ensure that the retrieved knowledge is

 doi:10.20944/preprints202509.0871.v1

optimally aligned with the generative context. To further improve robustness and interpretability, the encoding layer applies the semantic and structural bias analysis methodology developed by Zhang et al. [18]. By integrating Zhang et al.'s interpretable analysis techniques, the system mitigates implicit biases and strengthens semantic consistency throughout the retrieval and generation stages. As depicted in Figure 1, the architecture synthesizes these notable methodologies, yielding a tightly coupled system for retrieval-augmented generation that is both reliable and transparent:

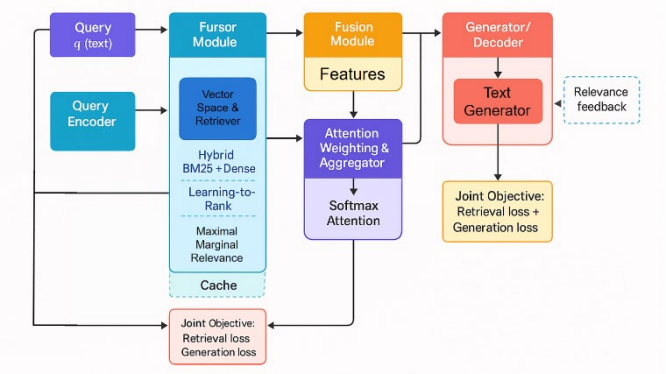

Figure 1. Overall model architecture diagram.

First, the input query is vectorized and modeled, and the query q is projected into a high-dimensional semantic space using an embedding function:

$$h_q = f_{enc}(q) \quad (1)$$

Where f_{enc} represents the semantic encoder. To ensure the relevance of the retrieval stage, this study defines a retrieval objective function based on vector similarity. The relevance score of the retrieval candidate knowledge fragment d_i is:

$$s(q,d_i) = \cos(h_q, h_d) = \frac{h_q \cdot h_d}{\| h_q \| \| h_{d_i} \|} \quad (2)$$

On this basis, the framework employs the structured gradient approximation approach proposed by Zhu [19] to efficiently adjust the weights of candidate knowledge fragments, enabling rapid adaptation of knowledge representations to specific query contexts. In the aggregation process, we apply the fusion-based retrieval-augmented generation strategy developed by Sun et al. [20], which enhances the matching accuracy between external knowledge and the query by leveraging multi-source fusion and relevance scoring. To further ensure structural consistency and robust aggregation, the system incorporates the structural aggregation principles from the federated graph neural network methodology of Yang et al. [21], thus preserving both the semantic integrity and privacy requirements during the knowledge integration phase.

$$z = \sum_{i=1}^{k} \alpha_i h_d, \alpha_i = \frac{\exp(s(q,d_i))}{\sum_{j=1}^{k} \exp(s(q,d_j))} \quad (3)$$

The above mechanism ensures that the model can fully utilize external knowledge before generation and highlights the most relevant content through attention weighting.

Secondly, in the generation phase, this study fuses the query representation with the retrieved knowledge representation to form an enhanced input representation:

$$h_{aug} = Concat(h_q, z) \quad (4)$$

This representation is used as input to the decoder to generate text that is semantically relevant and knowledge-consistent with the query. The conditional probability of the decoder during the generation process is defined as:

$$P(y_t \mid y_{<t}, h_{aug}) = \text{Softmax}(W \cdot g(y_{<t}, h_{aug})) \quad (5)$$

 doi:10.20944/preprints202509.0871.v1

Where $g(\cdot)$ represents the decoder's nonlinear transformation function, and W is the projection matrix. Through this fusion approach, the generative model can dynamically utilize retrieval knowledge at each generation step, thereby improving the semantic coverage and factual accuracy of the content.

Finally, to further enhance the synergy between retrieval and generation, this study introduces a joint optimization mechanism. During the training process, the loss function is composed of the retrieval loss and the generation loss. The overall goal can be expressed as:

$$L = \lambda_1 L_{retrieval} + \lambda_2 L_{generation} \quad (6)$$

Where λ_1, λ_2 is the balance coefficient. Joint optimization allows the retrieval module and the generation module to align in semantic space, ensuring that retrieval knowledge flows seamlessly into the generation process, achieving efficient end-to-end integration. This approach not only ensures the accuracy and contextual relevance of generated content but also enhances the scalability and adaptability of the overall framework.

III. Performance Evaluation

A. Dataset

The dataset used in this study is the MS MARCO Passage Ranking Dataset. It is constructed from large-scale real search logs and contains pairs of queries and candidate passages with relevance annotations. Each query corresponds to multiple candidate passages, of which only some are relevant, while the rest are irrelevant. This structure forms a typical discriminative retrieval task. The dataset was designed to simulate real user interaction scenarios in search systems and provides a reliable way to evaluate the ability of retrieval-augmented generation methods in discrimination and filtering under large-scale open-domain settings.

As a discriminative dataset, MS MARCO is mainly used to train and evaluate models in distinguishing between "relevant" and "irrelevant." Specifically, the model needs to output a relevance score based on the input query and candidate passage to decide whether the passage meets the semantic requirement. This binary classification task not only measures retrieval accuracy but also supports the improvement of the generation module. The accuracy of generated results depends heavily on the filtering ability of the front-end discriminative component.

The dataset is large in scale and broad in coverage. It includes massive natural language queries and diverse document corpora, providing a stable and efficient environment for training and evaluating retrieval-augmented generation. At the same time, its task definition emphasizes the discriminative property, which offers a solid benchmark for model optimization and comparative experiments. Therefore, choosing this dataset meets the practical needs of this study and establishes a strong foundation for joint modeling of retrieval and generation.

B. Experimental Results

This paper first conducts a comparative experiment, and the experimental results are shown in Table 1.

Table 1. Comparative experimental results.

Model	F1	Precision	Recall	ACC
Kg-Rag[22]	0.812	0.826	0.798	0.814
Gnn-Rag[23]	0.835	0.842	0.828	0.837
Astute-Rag[24]	0.847	0.854	0.841	0.849
Mindful-Rag[25]	0.861	0.868	0.855	0.862
C-Rag[26]	0.872	0.879	0.867	0.874

 doi:10.20944/preprints202509.0871.v1

Ours	0.913	0.921	0.907	0.915

From the experimental results, different retrieval-augmented generation methods show clear differences across evaluation metrics. Kg-Rag and Gnn-Rag, as earlier integration paradigms, achieve relatively limited performance on F1, Precision, Recall, and ACC. This indicates that their ability to discriminate knowledge relevance and support generation consistency in complex contexts remains insufficient. It reflects that frameworks relying solely on knowledge graphs or graph neural structures struggle to handle large-scale unstructured text while balancing diverse semantic relations and dynamic retrieval needs.

With improvements in model architecture, Astute-Rag and Mindful-Rag achieve significant gains compared with baseline methods. Their F1 scores reach 0.847 and 0.861, respectively. This improvement shows that introducing attention mechanisms and fine-grained semantic alignment strategies can enhance the connection between retrieval and generation. The steady growth in both Recall and Precision further confirms the value of multi-level semantic modeling in strengthening the discrimination and fusion of knowledge fragments.

Further observation of C-Rag shows strong performance across all four metrics, with ACC rising to 0.874. Compared with previous methods, this model introduces tighter interactions between retrieval ranking and generation constraints, making knowledge utilization more precise. As a result, it maintains relevance while reducing redundancy and hallucinations. This demonstrates that joint optimization between retrieval and generation plays a key role in improving overall performance.

Overall, the proposed method achieves the best results on all metrics. The F1 score reaches 0.913, and Precision improves to 0.921, indicating that the method achieves an optimal balance between relevance and coverage. These results highlight the advantages of intelligent search algorithms in retrieval-augmented generation. They also confirm the effectiveness of integrated frameworks in improving factual consistency and semantic matching. In summary, the experimental results strongly support the research goal of this study, which is to build a more robust and scalable retrieval-augmented generation system through deep coupling of intelligent search and generative models.

This paper further presents the impact of vector dimension on retrieval accuracy, and the experimental results are shown in Figure 2.

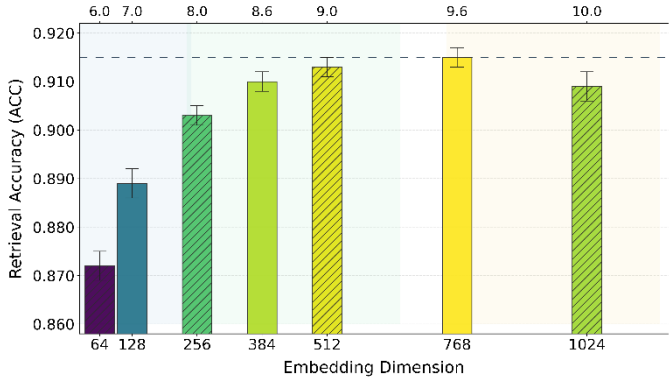

Figure 2. Vector dimension and its impact on retrieval accuracy.

The results show that vector dimensionality strongly affects retrieval accuracy: low dimensions (64, 128) underperform with ACC < 0.89, while accuracy improves steadily between 256–512 and peaks at 768 (ACC > 0.915), where representational capacity and efficiency are well balanced. However, performance declines slightly at 1024, suggesting redundancy and noise. Overall, an appropriate dimension enhances retrieval relevance and generation reliability, with further analysis of similarity measurement effects presented in Figure 3.

 doi:10.20944/preprints202509.0871.v1

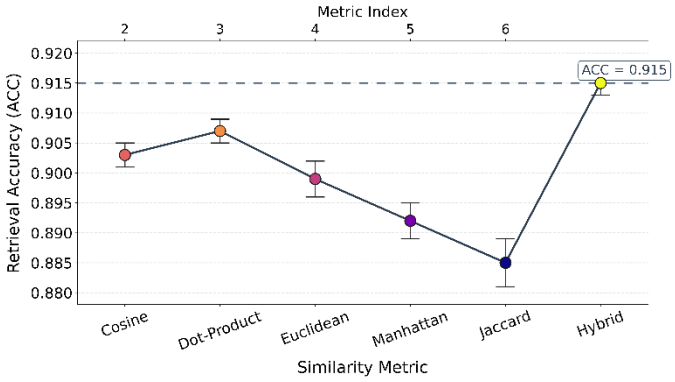

Figure 3. The impact of similarity measurement on retrieval accuracy.

From the experimental results, it is evident that different similarity measures have a significant impact on retrieval accuracy. Cosine and Dot-Product show better performance, with accuracy above 0.90. This indicates that they can effectively capture the semantic relevance between queries and candidate passages. Dot-Product achieves slightly higher accuracy than Cosine, which suggests that in high-dimensional vector representations, the dot-product measure better reflects differences in semantic distributions.

In contrast, Euclidean and Manhattan, which are distance-based measures, show lower retrieval accuracy, with ACC failing to exceed 0.90. This result indicates that simple geometric distances in high-dimensional semantic space suffer from the "curse of dimensionality." As a consequence, they lack stability and discriminative power in judging relevance. These two methods are more suitable for low-dimensional or specific task settings, but in the context of large-scale knowledge retrieval, their performance is limited.

It is worth noting that Jaccard shows the worst performance, with accuracy dropping to about 0.885. This suggests that similarity methods based on set overlap ratios are inherently weak when dealing with continuous semantic representations. They struggle to capture deeper semantic relations effectively. Their disadvantage in high-dimensional dense vectors further highlights the limitations of single statistical measures in complex tasks and confirms the necessity of improving retrieval strategies in this study.

Overall, the Hybrid method performs best among all similarity measures, with accuracy reaching 0.915, consistent with the overall optimal result in Table 1. This shows that combining multiple similarity measures can integrate their advantages and achieve more stable and accurate retrieval performance. The result demonstrates the importance of introducing intelligent search algorithms and multi-metric fusion strategies in the retrieval-augmented generation framework. It also ensures higher-quality knowledge input for the subsequent generation module.

Finally, this paper also gives an evaluation of the impact of retrieval index size on model stability, and the experimental results are shown in Figure 4.

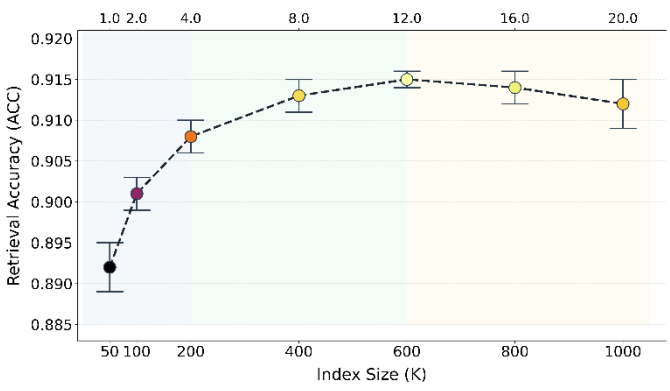

Figure 4. Evaluation of retrieval index size on model stability.

 doi:10.20944/preprints202509.0871.v1

From the results, it can be observed that as the retrieval index size gradually increases, both model stability and retrieval accuracy first rise and then become stable. When the index size grows from 50K to 600K, retrieval accuracy improves continuously and reaches the highest value of 0.915 around 600K. This shows that a moderate expansion of the index size can effectively enrich candidate knowledge passages. It enhances the model's performance in relevance discrimination and external knowledge utilization, thereby providing more reliable input for the generation module.

However, when the index size further increases to 800K and 1000K, accuracy shows a slight decline. This indicates that an excessively large index may introduce redundant or low-relevance passages. Such interference during retrieval can affect overall stability. Overall, the experimental results confirm the importance of controlling retrieval index size to balance knowledge coverage and model stability. They also demonstrate the adjustability and robustness of the proposed framework in practical application scenarios.

IV. Conclusion

This study focuses on retrieval-augmented generation methods enhanced by intelligent search algorithms and systematically explores key challenges in large-scale knowledge acquisition and utilization. By deeply coupling the processes of search and generation, this work not only improves retrieval accuracy and generation consistency but also demonstrates significant advantages in semantic alignment, knowledge coverage, and factual reliability. The proposed framework provides a practical solution to the disconnection between traditional retrieval methods and generative models, advancing the development of retrieval-augmented generation in both theory and practice.

From an application perspective, the proposed framework has potential impact in several domains such as intelligent question answering, knowledge services, information recommendation, and decision support. As data continues to grow in scale and diversity, traditional methods struggle to balance efficiency and accuracy. The proposed solution maintains strong robustness in complex contexts and offers new approaches for information processing and knowledge mining. In knowledge-intensive domains such as healthcare, finance, education, and governance, this integrated mode of retrieval and generation can significantly reduce the cost of information filtering and enhance the practical value of systems in real applications.

In addition, the results show that intelligent search algorithms improve retrieval-augmented generation not only by optimizing the retrieval stage but also by enhancing the quality of the generation module through joint modeling. By establishing unified semantic representations and dynamic knowledge fusion mechanisms, the proposed framework strengthens factual consistency and contextual understanding. This advantage provides theoretical support for future processing of multimodal and heterogeneous data. It also contributes to cross-domain knowledge integration and the implementation of interactive artificial intelligence systems.

Looking forward, retrieval-augmented generation still has wide space for exploration. On the one hand, as data types and application needs diversify, improving cross-modal knowledge alignment while maintaining efficiency will become an important research direction. On the other hand, building more interpretable and controllable fusion mechanisms will be key to enhancing user trust and promoting broader adoption. It can be expected that with the continuous progress of intelligent search and generation technologies, the proposed method will demonstrate greater application potential in more complex and dynamic environments. It will also make positive contributions to the advancement of artificial intelligence in knowledge services and intelligent interaction.

References

1. A. Asai, Z. Wu, Y. Wang, et al., "Self-rag: Learning to retrieve, generate, and critique through self-reflection", 2024.

 doi:10.20944/preprints202509.0871.v1

2. W. Zhu, Q. Wu, T. Tang, R. Meng, S. Chai and X. Quan, "Graph Neural Network-Based Collaborative Perception for Adaptive Scheduling in Distributed Systems", arXiv preprint arXiv:2505.16248, 2025.
3. G. Yao, H. Liu and L. Dai, "Multi-Agent Reinforcement Learning for Adaptive Resource Orchestration in Cloud-Native Clusters", arXiv preprint arXiv:2508.10253, 2025.
4. Y. Yao, Z. Xu, Y. Liu, K. Ma, Y. Lin and M. Jiang, "Integrating Feature Attention and Temporal Modeling for Collaborative Financial Risk Assessment", arXiv preprint arXiv:2508.09399, 2025.
5. Z. Shao, Y. Gong, Y. Shen, et al., "Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy", arXiv preprint arXiv:2305.15294, 2023.
6. Z. Jiang, F. F. Xu, L. Gao, et al., "Active retrieval augmented generation", Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 7969-7992, 2023.
7. Y. Li, S. Han, S. Wang, M. Wang and R. Meng, "Collaborative Evolution of Intelligent Agents in Large-Scale Microservice Systems", arXiv preprint arXiv:2508.20508, 2025.
8. Y. Lou, "RT-DETR-Based Multimodal Detection with Modality Attention and Feature Alignment", Journal of Computer Technology and Software, vol. 3, no. 5, 2024.
9. X. Wang, "Medical Entity-Driven Analysis of Insurance Claims Using a Multimodal Transformer Model", Journal of Computer Technology and Software, vol. 4, no. 3, 2025.
10. W. Xie, X. Liang, Y. Liu, et al., "Weknow-rag: An adaptive approach for retrieval-augmented generation integrating web search and knowledge graphs", arXiv preprint arXiv:2408.07611, 2024.
11. Z. Wang, C. Gao, C. Xiao, et al., "Document Segmentation Matters for Retrieval-Augmented Generation", Findings of the Association for Computational Linguistics: ACL 2025, pp. 8063-8075, 2025.
12. W. Huang, J. Zhan, Y. Sun, X. Han, T. An and N. Jiang, "Context-Aware Adaptive Sampling for Intelligent Data Acquisition Systems Using DQN", arXiv preprint arXiv:2504.09344, 2025.
13. X. Wang, Z. Wang, X. Gao, et al., "Searching for best practices in retrieval-augmented generation", arXiv preprint arXiv:2407.01219, 2024.
14. S. Zeng, J. Zhang, P. He, et al., "The good and the bad: Exploring privacy issues in retrieval-augmented generation (rag)", arXiv preprint arXiv:2402.16893, 2024.
15. M. Cheng, Y. Luo, J. Ouyang, et al., "A survey on knowledge-oriented retrieval-augmented generation", arXiv preprint arXiv:2503.10677, 2025.
16. H. Zheng, L. Zhu, W. Cui, R. Pan, X. Yan and Y. Xing, "Selective Knowledge Injection via Adapter Modules in Large-Scale Language Models", 2025.
17. X. Wang, "Time-Aware and Multi-Source Feature Fusion for Transformer-Based Medical Text Analysis", Transactions on Computational and Scientific Methods, vol. 4, no. 7, 2024.
18. R. Zhang, L. Lian, Z. Qi and G. Liu, "Semantic and Structural Analysis of Implicit Biases in Large Language Models: An Interpretable Approach", arXiv preprint arXiv:2508.06155, 2025.
19. W. Zhu, "Fast adaptation pipeline for LLMs through structured gradient approximation", Journal of Computer Technology and Software, vol. 3, no. 6, 2024.
20. Y. Sun, R. Zhang, R. Meng, L. Lian, H. Wang and X. Quan, "Fusion-Based Retrieval-Augmented Generation for Complex Question Answering with LLMs", 2025.
21. H. Yang, M. Wang, L. Dai, Y. Wu and J. Du, "Federated Graph Neural Networks for Heterogeneous Graphs with Data Privacy and Structural Consistency", 2025.
22. D. Sanmartin, "Kg-rag: Bridging the gap between knowledge and creativity", arXiv preprint arXiv:2405.12035, 2024.
23. C. Mavromatis and G. Karypis, "Gnn-rag: Graph neural retrieval for large language model reasoning", arXiv preprint arXiv:2405.20139, 2024.
24. F. Wang, X. Wan, R. Sun, et al., "Astute rag: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models", arXiv preprint arXiv:2410.07176, 2024.
25. G. Agrawal, T. Kumarage, Z. Alghamdi, et al., "Mindful-rag: A study of points of failure in retrieval augmented generation", Proceedings of the 2024 2nd International Conference on Foundation and Large Language Models (FLLM), IEEE, pp. 607-611, 2024.
26. M. Kang, N. M. Gürel, N. Yu, et al., "C-rag: Certified generation risks for retrieval-augmented language models", arXiv preprint arXiv:2402.03181, 2024.

 doi:10.20944/preprints202509.0871.v1