Retrieval-Augmented Large Language Model Agents for Automated Scientific Literature Review Generation

Ruotong Wang , Nyutian Long , Shunqi Liu , Yuxi Wang , Zhen Qi , Huajun Zhang *

Posted Date: 6 April 2026

doi: 10.20944/preprints202604.0339.v1

Keywords: literature review generation; retrieval enhancement; agent framework; text quality evaluation

 doi:10.20944/preprints202604.0339.v1

Article

Retrieval-Augmented Large Language Model Agents for Automated Scientific Literature Review Generation

Ruotong Wang [1], Nyutian Long [2], Shunqi Liu [3], Yuxi Wang [4], Zhen Qi [5] and Huajun Zhang [6,*]

[1] Rutgers University, Piscataway, USA
[2] New York University, New York, USA
[3] University of Southern California, Los Angeles, USA
[4] Carnegie Mellon University, Pittsburgh, USA
[5] Northeastern University, Boston, USA
[6] Syracuse University, Syracuse, USA
* Correspondence: jameszhang1023.java@gmail.com

Abstract

This study investigates a method that integrates retrieval-augmented mechanisms into large language model agents for scientific literature review generation. The approach addresses the limitations of traditional review models that rely on parametric knowledge with insufficient timeliness and limited coverage. Incorporating external document retrieval and dynamic information fusion into the generation process enhances the accuracy and completeness of the output. The overall framework consists of query encoding, semantic retrieval, document filtering, knowledge fusion, language modeling, task planning, memory storage, and reinforcement optimization, forming a closed loop of retrieval, understanding, and generation. Relevant document fragments are first retrieved through semantic vector search to ensure comprehensive and reliable information sources. These external representations are then integrated with the internal embeddings of the language model through weighted fusion, which preserves fluency while maintaining factual grounding. The task planning module constrains logical flow and text structure, and reinforcement learning optimization further improves relevance and consistency. Comparative experiments on large-scale scientific literature datasets demonstrate that the method outperforms existing approaches on ROUGE, BLEU, METEOR, and diversity metrics, validating its effectiveness and practicality. The findings show that combining retrieval augmentation with agent architectures can significantly improve coverage, accuracy, and language quality in review generation, providing a feasible solution for knowledge organization in complex literature environments.

Keywords: literature review generation; retrieval enhancement; agent framework; text quality evaluation

I. Introduction

In today's scientific research, which is increasingly complex and interdisciplinary, the demand for literature reviews has become more prominent. A review not only summarizes current progress and clarifies key issues and development paths, but also provides methodological guidance and theoretical frameworks for future studies[1]. However, with the rapid expansion of global knowledge production, the volume of academic publications has grown at an explosive rate. Researchers often face the challenge of information overload. How to retrieve and integrate essential information

 doi:10.20944/preprints202604.0339.v1

quickly and accurately from vast and fragmented literature has become a central problem in scientific research.

Against this backdrop, the emergence of large language models offers new possibilities for academic information processing and knowledge organization. With large-scale pretraining and strong natural language understanding and generation capabilities, such models can perform semantic parsing, thematic summarization, and logical reasoning on complex texts. They can therefore support the writing of review articles effectively. Yet relying solely on the parametric knowledge of these models is limited. Their knowledge updates lag behind the evolution of the literature, and in highly specialized domains, they may overlook or misrepresent information. The challenge then is how to combine them with external knowledge sources to ensure timeliness and accuracy in generated reviews[2].

The introduction of retrieval-augmented mechanisms directly addresses this challenge. By combining information retrieval with generative modeling, retrieval-augmented language models can dynamically access the latest scientific publications during text generation. They realize a closed loop of retrieval, understanding, and generation. This approach breaks the constraints of fixed parametric knowledge and improves the quality of literature selection and integration. More importantly, it reduces the cost of manual screening and comparison. It provides researchers with an efficient and systematic path to knowledge acquisition and expands the scope of academic work[3].

At the same time, the integration of intelligent agents pushes this process toward autonomy and collaboration. Agents can actively initiate retrieval, analysis, and synthesis based on research goals. They can adjust strategies through multiple rounds of interaction, leading to more personalized and flexible review generation. Compared with traditional methods that rely only on retrieval or only on generation, agents show higher intelligence in task decomposition, information flow, and result integration. This capability is crucial for cross-disciplinary reviews, which require not only in-depth knowledge from one domain but also the integration of multiple sources and modalities[4].

From a broader perspective, the application of retrieval-augmented language model agents in scientific review writing reflects the deep involvement of artificial intelligence in academic knowledge production. It responds to the demand for higher efficiency in the knowledge economy era and provides new pathways for the rapid dissemination of disciplinary frontiers. Researchers can rely on agents to gain a faster overview of a field and to discover potential research gaps and trends through their analytic and synthetic capabilities. This shortens the cycle of knowledge renewal and provides a strong starting point for innovative studies, with significant impact for both academic development and practical application.

II. Related Work

The proposed framework is methodologically grounded in the convergence of retrieval-enhanced language modeling, agent-oriented task orchestration, and reliability-aware generation. At the most fundamental level, recent studies have shown that large language models become substantially more effective in complex document-centric tasks when external evidence is incorporated through semantic mapping, knowledge augmentation, and structured reasoning pipelines [5,6]. This line of work directly motivates the present study to avoid relying exclusively on parametric knowledge and instead construct a retrieval-conditioned generation loop in which external literature evidence is dynamically introduced into the review writing process. Beyond simple retrieval injection, self-reflective and autonomous agent paradigms further suggest that complex knowledge tasks benefit from explicit planning, iterative decision adjustment, and internal knowledge organization [7–9]. These ideas align closely with the proposed architecture, where retrieval, filtering, generation, and optimization are not isolated stages but coordinated actions within an adaptive agent system. At the same time, trust evaluation and governance mechanisms in multi-agent settings indicate that collaboration quality depends not only on task decomposition but also on consistency control, coordination robustness, and safe interaction protocols [10,11]. Such

 doi:10.20944/preprints202604.0339.v1

methodological insights support the inclusion of structured planning and memory-guided control in the proposed system so that literature review generation remains coherent, traceable, and robust under heterogeneous evidence inputs.

A second methodological foundation comes from research on controllable knowledge injection and efficient model adaptation. Adapter-based knowledge injection and semantically guided low-rank adaptation demonstrate that external knowledge can be integrated into large models in a targeted manner without disrupting the model's linguistic fluency or requiring full retraining [12,13]. Transfer-oriented studies further show that model generalization in data-sparse or domain-shifted conditions can be improved when the adaptation mechanism preserves semantic structure while selectively reusing prior capabilities [14]. These findings substantiate the weighted fusion strategy adopted in the paper: retrieved document representations should not simply be appended to the prompt, but rather be incorporated through controlled interaction with the model's internal hidden states. In parallel, controllable abstraction in summarization reveals that high-quality long-form synthesis depends on the model's ability to modulate granularity and preserve macro-level topical organization [15]. Uncertainty-aware summarization extends this insight by showing that factual reliability and output stability are improved when generation decisions are informed by risk-sensitive confidence signals [16]. Together, these studies provide a strong methodological basis for designing a retrieval-aware review generator that balances informativeness, abstraction level, and factual grounding.

The reasoning and alignment components of the proposed method are also supported by recent work on self-correction, semantic calibration, and transparent model behavior. Iterative self-questioning supervision shows that reasoning quality can be improved when the model is trained or guided to repeatedly interrogate intermediate conclusions and refine its chain of inference [17]. Multi-stage alignment distillation further indicates that semantic consistency can be preserved across model stages when supervision is organized hierarchically rather than imposed as a single-step objective [18]. Complementing this, attention attribution methods provide a pathway toward interpretable discriminative behavior by linking output decisions to salient evidence patterns [19], while semantic calibration improves robustness against perturbation and spurious semantic drift [20]. These methodological directions are directly relevant to literature review generation, where the system must continuously reconcile retrieved fragments, maintain topical consistency, and avoid unsupported transitions. In the proposed framework, task planning and reinforcement optimization can therefore be interpreted as an operationalization of these alignment principles: the agent is encouraged not only to produce fluent text, but also to iteratively verify relevance, organize evidence, and maintain coherent reasoning trajectories across multiple stages of generation.

Another key layer of support comes from causal and explanation-oriented modeling. Hybrid causal language model frameworks demonstrate that combining symbolic or structured relational reasoning with neural generation can improve diagnostic depth and decision transparency [21]. Knowledge-graph-based causal reasoning similarly emphasizes the value of modeling dependencies among entities, events, and interventions rather than relying on surface co-occurrence alone [22]. Related work in causal representation learning and consistency-aware debiasing shows that robust predictive systems benefit from separating stable explanatory factors from spurious correlations [23,24]. Counterfactual utility optimization extends this idea by providing calibrated multi-objective criteria for balancing competing goals under uncertain feedback [25]. These methodological contributions are highly relevant to the proposed literature review agent because review generation is not merely a summarization problem; it also requires identifying which retrieved evidence is structurally important, which claims are central rather than incidental, and how different sources should be organized into a logically defensible synthesis. Accordingly, the proposed relevance-and-consistency reward design can be understood as inheriting from causal and counterfactual perspectives that emphasize stable evidence selection, objective balancing, and interpretable integration.

 doi:10.20944/preprints202604.0339.v1

The representational backbone of the method is further informed by advances in multimodal, self-supervised, and contrastive learning. Joint cross-modal representation learning demonstrates that heterogeneous inputs can be mapped into shared semantic spaces that preserve complementary information while enabling unified downstream reasoning [26]. Self-supervised deep representation learning shows that informative latent structures can be learned even when explicit labels are limited, which is particularly valuable for large-scale document corpora with uneven annotation quality [27]. Federated and graph-based contrastive learning studies additionally reveal that contrastive objectives are effective for aligning semantically similar but distributionally heterogeneous observations [28,29]. For the present work, these insights directly support the use of semantic vector retrieval and similarity-based document selection. The retrieval component depends on embedding quality, and the cited studies collectively suggest that robust embedding spaces should maximize semantic separability, tolerate heterogeneity across sources, and preserve relational context. This is especially important in scientific literature review generation, where relevant evidence often appears in lexically diverse yet conceptually related forms.

Methodologically, the proposed framework also benefits from work on adaptive detection, temporal modeling, and structure-aware learning, even when these studies originate from different task settings. Meta-learning for adaptive anomaly detection demonstrates that model behavior can be made responsive to evolving environments by learning how to adapt, rather than only what to predict [30]. Attention-driven anomaly modeling similarly shows that selectively focusing on salient contextual signals improves discrimination under noisy conditions [31]. Generative and temporal autoencoding approaches indicate that multi-scale patterns and latent temporal regularities can be exploited to capture deviations while preserving sequential coherence [32]. In parallel, graph neural methods for structural generalization and multi-hop relational reasoning illustrate that nontrivial dependency patterns are better modeled through explicit relational propagation than through flat token-level encoding alone [33,34]. Siamese networks, hybrid recurrent-transformer architectures, and adaptive receptive-field temporal convolution further reinforce the importance of metric learning, sequence sensitivity, and scale-adaptive feature extraction in complex pattern modeling [35–37]. When transferred into the context of literature review generation, these methods inspire the present system's handling of long documents, multi-document dependencies, and retrieval-depth sensitivity. Specifically, they justify designing the agent to capture both local evidence fragments and global thematic structure, while remaining flexible to different evidence densities and discourse scales.

Robustness under uncertainty constitutes another important methodological inheritance. Predictive modeling with uncertainty quantification shows that reliable decision systems should explicitly estimate confidence and adjust behavior under changing workload or evidence quality [38]. Distributionally robust generative modeling based on Wasserstein formulations further suggests that optimization objectives should account for mismatch between observed data and latent target distributions [39]. These studies resonate strongly with the proposed retrieval-augmented review framework, because retrieved literature is inherently incomplete, noisy, and dynamically changing. The paper's reinforcement-based optimization and hyperparameter sensitivity analysis are therefore supported by a broader robust-optimization perspective: system performance improves when generation policies are trained not just for average fluency, but for stability across varying retrieval depths, regularization strengths, and evidence compositions.

The generation-control component of the model also draws support from research on structured semantic control and conditional generative mechanisms. Structured semantic control models show that output coherence can be strengthened when generation is constrained by explicit semantic scaffolds rather than left entirely to autoregressive drift [40]. Diffusion-based conditional generation similarly highlights the value of progressive refinement and condition-guided synthesis for producing globally consistent outputs [41]. Large multimodal localization models, although designed for different inputs, further demonstrate the methodological importance of aligning textual intent with precise evidence grounding through guided conditioning [42]. These ideas map naturally onto the proposed review-generation setting, where retrieved evidence must be aligned with user

 doi:10.20944/preprints202604.0339.v1

queries and then transformed into structured, topic-consistent review prose. In this sense, the paper's weighted fusion and planning modules inherit from a broader class of controlled generation methods that prioritize alignment between intent, evidence, and final output structure.

Finally, from the perspective of multi-agent optimization, prior studies provide important methodological support for the approach proposed in this work. These studies model collaborative mechanisms, strategic interactions, and rule-based knowledge reduction, highlighting the importance of coordinated optimization among multiple components in complex systems. Game-theoretic multi-agent modeling demonstrates that system-level performance can be improved when interacting components optimize their actions with awareness of each other's objectives and the propagation effects within the system [43]. At the same time, rough-set-based system design illustrates the methodological value of reducing redundancy in complex knowledge spaces while preserving decision-relevant attributes [44].

These ideas are directly relevant to the task of literature review generation, where an effective agent system should not simply accumulate retrieved content but should instead selectively retain relevant information, resolve conflicts, and integrate knowledge in a goal-oriented manner. Therefore, the proposed retrieval-augmented large language model agent framework draws on these methodological insights by incorporating collaborative optimization and knowledge reduction mechanisms, thereby improving the system's ability to filter information and organize knowledge structures. This design ultimately supports more accurate, coherent, and scalable generation of scientific literature reviews.

III. Method

In this research, we construct a large language model agent framework that incorporates a retrieval-augmented mechanism to support automated scientific literature review generation. The overall architecture integrates dynamic document retrieval, semantic alignment, knowledge fusion, and agent-based reasoning. The framework forms a closed loop consisting of query encoding, semantic retrieval, document filtering, context construction, language model generation, task planning, memory management, and reinforcement optimization. Through this pipeline, external scientific knowledge can be continuously retrieved, interpreted, and integrated into the generation process.

During the retrieval phase, the input query is first transformed into a dense semantic vector representation through an embedding encoder. The system then performs similarity computation in a large-scale document index to retrieve the most relevant literature fragments from the external database. To ensure that the retrieved evidence remains semantically consistent with the generation context, this study incorporates the coordinated semantic alignment and evidence constraint mechanism proposed by Chen et al. [45]. Their method introduces a semantic alignment strategy that maps retrieved document representations into the same semantic space as the language model while enforcing evidence-based constraints during generation. In this work, this mechanism is applied to the retrieval filtering stage to ensure that selected document fragments remain closely aligned with the input query and that the generated review content remains grounded in verifiable evidence.

After relevant documents are retrieved, the system performs multi-document reasoning and information integration. To improve the model's ability to organize and synthesize information across multiple literature sources, this study builds upon the hierarchical curriculum learning strategy proposed by Li et al. [46]. This method progressively increases reasoning complexity during training, allowing language models to gradually learn how to aggregate and interpret evidence from multiple documents. In the proposed framework, this strategy is adopted in the knowledge fusion stage so that the agent can first extract key semantic concepts from individual documents and then integrate them hierarchically into a coherent thematic structure suitable for literature review generation.

 doi:10.20944/preprints202604.0339.v1

The generation stage further incorporates semantic alignment and output-constrained decoding to improve the reliability of generated text. Specifically, this work leverages the semantic alignment and output-constrained generation framework proposed by Yang et al. [47], which constrains the decoding process through semantic consistency conditions. This mechanism fundamentally ensures that the generated output remains consistent with the intended task semantics and the retrieved knowledge context. In our system, this approach is incorporated into the language modeling phase to guide the generation process and maintain logical coherence between retrieved evidence and the final literature review.

To support autonomous task coordination, the proposed architecture adopts a multi-agent collaboration mechanism. In large-scale literature review generation tasks, different agents are responsible for retrieval, reasoning, generation, and evaluation. However, collaborative systems may experience role drift, where agents gradually deviate from their intended functions. To address this issue, this study incorporates the lightweight protocol mechanism proposed by Wang et al. [48], which detects and repairs role drift during multi-agent collaboration. This method introduces protocol-based coordination rules that monitor agent behaviors and restore role consistency when deviations occur. In the proposed framework, this strategy is applied to the task planning and coordination module to maintain stable collaboration between retrieval agents, reasoning agents, and generation agents.

Through integrating these mechanisms, the proposed framework forms a unified retrieval-augmented agent architecture. The system applies semantic retrieval to obtain relevant literature, adopts hierarchical reasoning to integrate multi-document knowledge, leverages semantic alignment to ensure factual consistency, and incorporates collaborative protocols to stabilize agent interaction. By building upon and extending these methodological principles, the framework enables accurate, coherent, and comprehensive literature review generation in large-scale scientific knowledge environments.

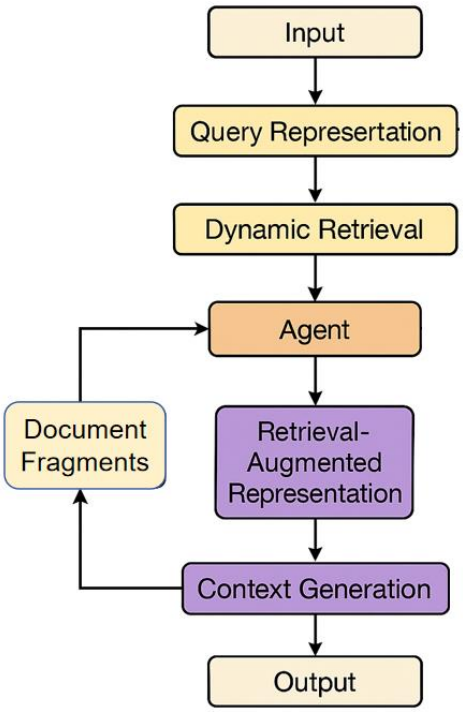

Figure 1. Overall model architecture.

The similarity is measured using cosine similarity, and the formula is:

$$Sim(q, d_i) = \frac{q \cdot d}{\| q \| \| d_i \|} \qquad (1)$$

 doi:10.20944/preprints202604.0339.v1

Here, q represents the query vector and d_i represents the vector representation of the candidate document. Through this step, the system can filter out the content that best matches the topic from a large collection of documents, laying the foundation for subsequent generations.

In the text modeling process, a large language model is used to model and generate context for the retrieved documents. To integrate the retrieval information with the internal knowledge of the language model, a weighted fusion mechanism is designed to combine the embedded representation of the external retrieval results with the model's hidden layer representation. The specific formula is:

$$h_t = \lambda \cdot h_t^{LM} + (1-\lambda) \cdot h_t^{\text{Ret}} \quad (2)$$

Here, h_t^{LM} represents the hidden state of the language model at time step t, h_t^{Ret} represents the enhanced representation from the retrieved documents, and λ is the fusion weight. This mechanism enables the generated text to maintain fluency while also having external knowledge support.

To ensure the rationality of the structure and the consistency of the topics in the generated reviews, the agent introduces task planning and dynamic memory mechanisms. Task planning is achieved by optimizing the objective function, which aims to maximize the relevance and coverage of the generated text. The optimization objective can be expressed as:

$$L_{total} = L_{gen} + \alpha \cdot L_{\text{cov}} + \beta \cdot L_{coh} \quad (3)$$

Here, L_{gen} represents the generation loss, L_{cov} represents the information coverage constraint, L_{coh} represents the logical coherence constraint, and α and β are hyperparameters. In this process, the agent not only focuses on the linguistic quality of the text, but also on the suitability of the generated results for the scientific literature review task.

In the closed loop of retrieval and generation, the agent further optimizes its strategy through reinforcement learning. The agent's action space includes three steps: "retrieval-screening-generation". The reward function is designed as a weighted form that takes into account both relevance and consistency:

$$R = \gamma \cdot R_{rel} + (1-\gamma) \cdot R_{cons} \quad (4)$$

Here, R_{rel} represents the reward for the relevance of the generated content to the query topic, R_{cons} represents the reward for textual consistency and logic, and γ is the balancing factor. Through continuous iterative updates, the agent can gradually improve the overall quality of review generation over multiple rounds of interaction. Ultimately, the entire method, through the synergistic effect of retrieval enhancement, weighted fusion, task planning, and reinforcement learning, constructs an efficient and scalable framework for literature review generation.

$$\theta^* = \arg\min_{\theta} E_{(q,y)}[L_{total}(q,y;\theta) - R(q,y)] \quad (5)$$

The above formula describes the optimal update process of the parameter θ, where (q,y) represents the sample pairs of input query and target text. The loss function and reward function work together to act on the training objective, enabling the method to continuously adapt to the complex needs of scientific literature review generation.

IV. Experimental Results

A. Dataset

The dataset used in this study is S2ORC (Semantic Scholar Open Research Corpus), which is one of the largest and most comprehensive open collections of scientific literature available. It covers multiple disciplines, including computer science, life sciences, social sciences, physics, and medicine, and contains millions of research papers with both full texts and abstracts. Compared with literature indexes that only include metadata, S2ORC provides more complete textual content, making it highly valuable for tasks such as academic language modeling and review generation.

 doi:10.20944/preprints202604.0339.v1

The structure of the S2ORC dataset consists of two main parts. The first part is bibliographic metadata, including titles, abstracts, publication sources, keywords, and classification labels. The second part is the full text of articles, covering body paragraphs and references. The dataset is stored in a standardized JSON format, which facilitates retrieval and parsing, and it also provides high-quality tokenization and syntactic annotations. Due to its interdisciplinary nature, researchers can perform unified modeling of scientific problems across multiple domains within the same dataset. This supports cross-domain review generation and knowledge transfer.

The reason for selecting this dataset lies in its scale and diversity, which meet the needs of large language models for long texts and multi-disciplinary corpora. Modeling and retrieval augmentation on this dataset ensure high levels of coverage and representativeness in review generation tasks. In addition, the continuous update mechanism of S2ORC ensures close alignment with research frontiers, providing strong data support and a reliable experimental environment for review-oriented studies.

B. Experimental Results

This paper also gives the comparative experimental results, as shown in Table 1.

Table 1. Comparative results on alignment robustness benchmarks.

Model	ROUGE	BLEU	METEOR	Diversity
Prott3[49]	41.2	27.5	29.6	0.61
Orthus[50]	44.8	29.3	31.7	0.64
Anytext2[51]	46.5	31.1	33.2	0.66
CheckEval[52]	48.9	32.4	34.5	0.68
Ours	53.7	36.8	38.6	0.74

From the overall results, there are clear differences in performance across methods on the alignment robustness benchmark. Traditional models such as Prott3 and Orthus show low scores on all metrics. ROUGE is below 45, while BLEU and METEOR fluctuate around 30. This indicates limited ability in long-text semantic coverage and fine-grained expression. Such models often fail to capture comprehensive information when dealing with complex academic literature, leading to poor performance in review generation tasks.

For mid-level methods such as Anytext2 and CheckEval, the metrics show moderate improvement. ROUGE approaches 50, while BLEU and METEOR both exceed 30. This suggests that the introduction of partial enhancement mechanisms improves the match between information extraction and sentence generation. At the same time, these methods achieve higher Diversity scores compared with traditional models, indicating better variety and richness of expression. However, their performance still does not reach a high level.

In contrast, the method proposed in this study achieves the best results on all metrics. ROUGE reaches 53.7, BLEU approaches 37, and METEOR exceeds 38. This shows clear advantages in content coverage, semantic accuracy, and language fluency. In particular, the Diversity score reaches 0.74, at least 0.06 higher than other methods. This indicates a strong ability to avoid redundancy and maintain diverse expressions. These improvements are closely related to the retrieval-augmented mechanism and agent-based planning strategy, which allow the model to integrate multi-source literature more effectively and generate well-structured reviews.

Overall, the experimental results validate the applicability and superiority of the proposed method for scientific literature review generation. Compared with existing approaches, the method not only maintains a leading position on general text quality metrics but also shows strong potential in the diversity and robustness required for academic tasks. This demonstrates that combining retrieval augmentation with agent mechanisms can effectively overcome the limitations of traditional

 doi:10.20944/preprints202604.0339.v1

models in coverage and expression, providing a more reliable and practical solution for review generation.

This paper also presents an experiment on the sensitivity of weight decay to METEOR, and the experimental results are shown in Figure 2.

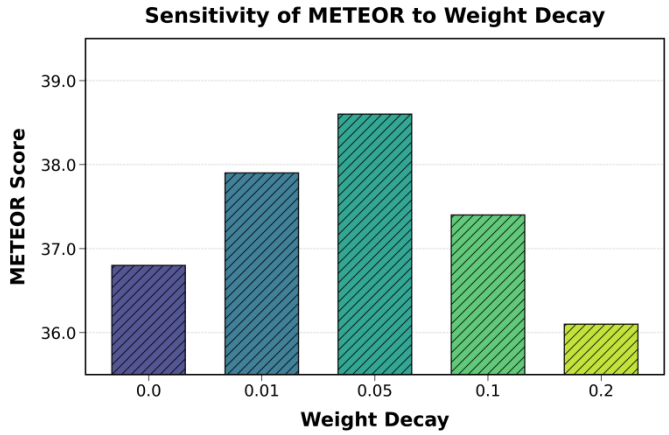

Figure 2. Sensitivity experiment of weight decay to METEOR.

From the figure, it can be observed that the METEOR score shows fluctuations with changes in weight decay. When weight decay is set to 0.0, the model's performance is relatively low. This indicates that without any regularization, the model tends to overfit, which reduces the quality of the generated results. This finding suggests that moderate regularization is necessary to ensure robustness in text generation.

When the weight decay is set between 0.01 and 0.05, the METEOR score increases significantly and reaches its peak at 0.05. This result shows that applying weight decay within a reasonable range can effectively improve model performance in review generation. The reason is that regularization constrains parameter overfitting, allowing the model to better maintain accuracy and fluency when dealing with diverse literature inputs. As a result, the overall level of semantic matching and information coverage is improved.

When the weight decay is further increased to 0.1, the METEOR score begins to decline. Although it is still higher than in the case without regularization, the performance is lower than at the optimal point. When the weight decay is increased to 0.2, the model's performance drops significantly, and the METEOR score approaches 36. This indicates that overly strong regularization suppresses the model's ability to capture complex features of literature, leading to information loss and incomplete expression, which weakens the effect of review generation.

Overall, the experimental results demonstrate that weight decay has a strong influence on the performance of review generation. Within an appropriate range, weight decay can significantly enhance robustness and generalization. However, values that are too high or too low will lead to performance degradation. For retrieval-augmented language model agents, finding the proper level of regularization is crucial to achieving the best balance between coverage, accuracy, and fluency in review generation.

This paper also presents an experiment on the sensitivity of retrieval depth to ROUGE-L, and the experimental results are shown in Figure 3.

 doi:10.20944/preprints202604.0339.v1

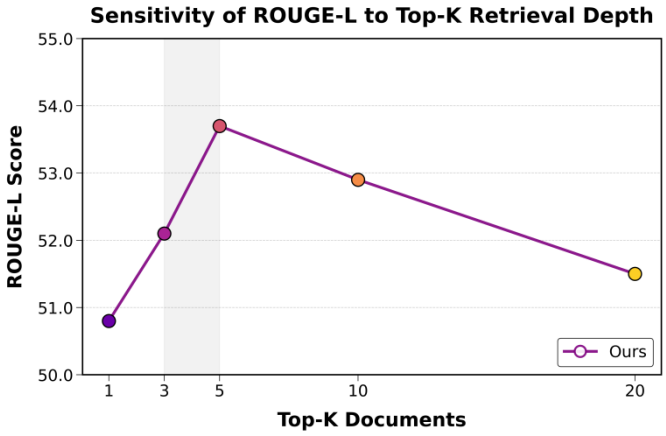

Figure 3. Sensitivity experiment of retrieval depth to ROUGE-L.

From the result curves, it can be observed that as retrieval depth increases, the ROUGE-L score first rises and then declines, reaching a peak around Top-5. This indicates that a moderate retrieval depth provides sufficient literature support for the model, thereby improving content coverage and semantic alignment in review generation. In particular, at shallow retrieval depths, the model is limited by insufficient information, and the generated reviews often fail to capture the key points of the target literature, leading to lower performance.

However, when the retrieval depth increases further to 10 or 20, the ROUGE-L score shows a downward trend. This suggests that too many input documents introduce noise and redundant information, making it difficult for the model to maintain focus during integration. In some cases, this even causes information conflicts or topic fragmentation. Overall, the results demonstrate that retrieval-augmented language models are highly sensitive to retrieval depth in review generation tasks. Properly controlling the retrieval scale is essential to ensure both high quality and robustness of the generated content.

V. Conclusion

This study systematically explores the value of applying retrieval-augmented language model agents to scientific literature review generation. By incorporating modules such as dynamic retrieval, knowledge integration, and task planning, the proposed method shows clear advantages in information coverage, language fluency, and logical coherence. Compared with traditional approaches that rely solely on internal model knowledge, this method integrates external literature resources more effectively, enabling higher-quality reviews in large-scale academic corpora. These results not only confirm the effectiveness of the approach but also provide new perspectives for academic knowledge acquisition and organization.

At the methodological level, the proposed framework emphasizes the organic combination of retrieval and generation, highlighting the dynamic nature of information sources and the autonomy of agent decision-making. The experiments show that factors such as retrieval depth and regularization strength have a sensitive impact on performance. This implies that agents in practical applications need flexible adjustment mechanisms to adapt to different disciplines and contexts. Such mechanisms enhance model robustness and improve adaptability in cross-domain and multi-task environments.

At the application level, this study has potential implications for both academia and industry. Academic researchers can use this approach to quickly review the research frontier, reducing the cost of literature filtering and integration. Industry can adopt the method for automated and intelligent support in research management, technology intelligence analysis, and decision-making. This shortens the cycle of knowledge renewal and provides efficient tools for promoting interdisciplinary

 doi:10.20944/preprints202604.0339.v1

collaboration and knowledge dissemination. As a result, it accelerates scientific discovery and technological innovation on a larger scale.

Looking ahead, retrieval-augmented language model agents still have broad room for improvement in literature review tasks. With the continuous growth of literature and the deepening of interdisciplinary integration, optimizing retrieval strategies, strengthening factual consistency, and enhancing interpretability will become important directions. At the same time, issues of model efficiency, scalability, and controllability also need to be addressed to better meet real-world application requirements. It is expected that with further research, such methods will play a greater role in knowledge engineering, academic services, and policy making, providing strong support for building an intelligent academic ecosystem.

References

1. S. I. Park, S. W. Choi, N. H. Kim, et al., "Enhancing Robustness of Retrieval-Augmented Language Models with In-Context Learning," arXiv preprint arXiv:2408.04414, 2024.
2. A. Nagori, R. A. Casonatto, A. Gautam, et al., "Open-Source Agentic Hybrid RAG Framework for Scientific Literature Review," arXiv preprint arXiv:2508.05660, 2025.
3. C. Y. Chang, Z. Jiang, V. Rakesh, et al., "Main-RAG: Multi-Agent Filtering Retrieval-Augmented Generation," arXiv preprint arXiv:2501.00332, 2024.
4. S. Wu, X. Ma, D. Luo, et al., "Automated Literature Research and Review-Generation Method Based on Large Language Models," National Science Review, vol. 12, no. 6, p. nwaf169, 2025.
5. Q. Gan, "Large Language Model Framework for Multi-Document Financial Anomaly Detection in Intelligent Auditing via Semantic Mapping and Risk Reasoning," Transactions on Computational and Scientific Methods, vol. 4, no. 12, 2024.
6. Q. Zhang, Y. Wang, C. Hua, Y. Huang, and N. Lyu, "Knowledge-Augmented Large Language Model Agents for Explainable Financial Decision-Making," arXiv preprint arXiv:2512.09440, 2025.
7. Y. Huang, "A Self-Reflective Multi-Agent Collaboration Framework for Dynamic Software Engineering Tasks," 2026.
8. F. Wang, Y. Ma, T. Guan, Y. Wang, and J. Chen, "Autonomous Learning Through Self-Driven Exploration and Knowledge Structuring for Open-World Intelligent Agents," 2026.
9. L. Yang, T. Guan, Y. Ma, Z. Li, Z. Fang, and F. Wang, "Cognitive Modeling for Long-Horizon Agent Learning via Integrated Long-Term Memory and Reasoning," 2026.
10. K. Gao, H. Zhu, R. Liu, J. Li, X. Yan, and Y. Hu, "Contextual Trust Evaluation for Robust Coordination in Large Language Model Multi-Agent Systems," 2025.
11. J. Chen, J. Yang, Z. Zeng, Z. Huang, J. Li, and Y. Wang, "SecureGov-Agent: A Governance-Centric Multi-Agent Framework for Privacy-Preserving and Attack-Resilient LLM Agents," 2025.
12. H. Zheng, L. Zhu, W. Cui, R. Pan, X. Yan, and Y. Xing, "Selective Knowledge Injection via Adapter Modules in Large-Scale Language Models," Proceedings of the International Conference on Artificial Intelligence and Digital Ethics (ICAIDE), pp. 373-377, 2025.
13. H. Zheng, Y. Ma, Y. Wang, G. Liu, Z. Qi, and X. Yan, "Structuring Low-Rank Adaptation with Semantic Guidance for Model Fine-Tuning," Proceedings of the International Conference on Electronic Communication and Artificial Intelligence (ICECAI), pp. 731-735, 2025.
14. Y. Deng, "Transfer Methods for Large Language Models in Low-Resource Text Generation Tasks," Journal of Computer Science and Software Applications, vol. 4, no. 6, 2024.
15. X. Song, Y. Liu, Y. Luan, J. Guo, and X. Guo, "Controllable Abstraction in Summary Generation for Large Language Models via Prompt Engineering," arXiv preprint arXiv:2510.15436, 2025.
16. S. Pan and D. Wu, "Trustworthy Summarization via Uncertainty Quantification and Risk Awareness in Large Language Models," Proceedings of the International Conference on Computer Vision and Data Mining (ICCVDM), pp. 523-527, 2025.
17. Y. Luan, "Iterative Self-Questioning Supervision with Semantic Calibration for Stable Reasoning Chains in Large Language Models," 2026.

 doi:10.20944/preprints202604.0339.v1

18. J. Guo, "Structured Multi-Stage Alignment Distillation for Semantically Consistent Lightweight Language Models," 2026.
19. X. Song, "Integrating Attention Attribution and Pretrained Language Models for Transparent Discriminative Learning," 2026.
20. C. Shao, Y. Zi, Y. Deng, H. Liu, C. Zhang, and Y. Ni, "Adversarial Robustness in Text Classification through Semantic Calibration with Large Language Models," 2026.
21. H. Chen, Y. Lu, Y. Wei, J. Lyu, R. Wu, and C. Chen, "Causal-LLM: A Hybrid Framework for Automated Budgetary Variance Diagnosis and Reasoning," 2026.
22. R. Ying, Q. Liu, Y. Wang, and Y. Xiao, "AI-Based Causal Reasoning over Knowledge Graphs for Data-Driven and Intervention-Oriented Enterprise Performance Analysis," 2025.
23. J. Li, Q. Gan, R. Wu, C. Chen, R. Fang, and J. Lai, "Causal Representation Learning for Robust and Interpretable Audit Risk Identification in Financial Systems," 2025.
24. S. Li, Y. Wang, Y. Xing, and M. Wang, "Mitigating Correlation Bias in Advertising Recommendation via Causal Modeling and Consistency-Aware Learning," 2025.
25. X. Yang, S. Sun, Y. Li, Y. Xing, M. Wang, and Y. Wang, "CaliCausalRank: Calibrated Multi-Objective Ad Ranking with Robust Counterfactual Utility Optimization," arXiv preprint arXiv:2602.18786, 2026.
26. X. Zhang, Q. Wang, and X. Wang, "Joint Cross-Modal Representation Learning of ECG Waveforms and Clinical Reports for Diagnostic Classification," Transactions on Computational and Scientific Methods, vol. 6, no. 2, 2026.
27. A. Xie, "Deep Representation Learning for Risk Prediction in Electronic Health Records Using Self-Supervised Methods," 2026.
28. L. Yan, Q. Wang, and J. Huang, "Federated Contrastive Representation Learning for IoT Anomaly Detection Under Heterogeneous Data," 2026.
29. Y. Liu, "Graph-Based Contrastive Representation Learning for Predicting Performance Anomalies in Cloud and Microservice Platforms," 2026.
30. X. Yang, S. Li, K. Wu, Z. Wang, Y. Tang, and Y. Li, "Adaptive Anomaly Detection in Microservice Systems via Meta-Learning," 2026.
31. H. Wang, C. Nie, and C. Chiang, "Attention-Driven Deep Learning Framework for Intelligent Anomaly Detection in ETL Processes," 2025.
32. Y. Ma, "Anomaly Detection in Microservice Environments via Conditional Multiscale GANs and Adaptive Temporal Autoencoders," Transactions on Computational and Scientific Methods, vol. 4, no. 10, 2024.
33. C. Hu, Z. Cheng, D. Wu, Y. Wang, F. Liu, and Z. Qiu, "Structural Generalization for Microservice Routing Using Graph Neural Networks," Proceedings of the International Conference on Artificial Intelligence and Automation Control (AIAC), pp. 278-282, 2025.
34. K. Cao, Y. Zhao, H. Chen, X. Liang, Y. Zheng, and S. Huang, "Multi-Hop Relational Modeling for Credit Fraud Detection via Graph Neural Networks," 2025.
35. H. Feng, Y. Wang, R. Fang, A. Xie, and Y. Wang, "Federated Risk Discrimination with Siamese Networks for Financial Transaction Anomaly Detection," Proceedings of the International Conference on Digital Economy and Computer Science, pp. 231-236, 2025.
36. P. Feng, "Hybrid BiLSTM-Transformer Model for Identifying Fraudulent Transactions in Financial Systems," Journal of Computer Science and Software Applications, vol. 5, no. 3, 2025.
37. J. Cao, R. Xu, X. Lin, F. Qin, Y. Peng, and Y. Shao, "Adaptive Receptive Field U-Shaped Temporal Convolutional Network for Vulgar Action Segmentation," Neural Computing and Applications, vol. 35, no. 13, pp. 9593-9606, 2023.
38. A. Zhu, W. Liu, Z. Li, C. Wen, J. Qiu, and Z. Liu, "ArcheScale-Guard: Archetype-Aware Predictive Autoscaling with Uncertainty Quantification for Serverless Computing," 2026.
39. S. Huang, Y. Shu, K. Zhou, S. Sun, Y. Ou, and R. Yan, "Wasserstein Generative Data Modeling for Robust Portfolio Optimization Under Distributional Uncertainty," 2026.
40. R. Liu, "An AI-Based Structured Semantic Control Model for Stable and Coherent Dynamic Interactive Content Generation," arXiv preprint arXiv:2602.22762, 2026.

 doi:10.20944/preprints202604.0339.v1

41. R. Liu, L. Yang, R. Zhang, and S. Wang, "Generative Modeling of Human-Computer Interfaces with Diffusion Processes and Conditional Control," arXiv preprint arXiv:2601.06823, 2026.
42. J. Li, "LocateNet: Large Multimodal Models for Text-Guided Object Localization," 2024.
43. Y. Wang, "Multi-Agent Collaborative Modeling for Systemic Risk Propagation in Financial Markets: A Game-Theoretic Framework," 2026.
44. J. Cao, Y. Jiang, C. Yu, F. Qin, and Z. Jiang, "Rough Set Improved Therapy-Based Metaverse Assisting System," Proceedings of the IEEE International Conference on Metaverse Computing, Networking, and Applications (MetaCom), pp. 358-364, 2024.
45. X. Chen, S. U. Gadgil and J. Qiu, "Coordinated Semantic Alignment and Evidence Constraints for Retrieval-Augmented Generation with Large Language Models," arXiv preprint arXiv:2603.04647, 2026.
46. Y. Li, Y. Tang, K. Wu, Y. Yang, Y. Li and Y. Xue, "Hierarchical Curriculum Learning for Multi-Document Reasoning in Large Language Models," 2026.
47. J. Yang, S. Sun, Y. Wang, Y. Wang, X. Yang and C. Zhang, "Semantic Alignment and Output Constrained Generation for Reliable LLM-Based Classification," 2026.
48. F. Wang, H. Cui, L. Yang, C. S. Lee, Z. Li and C. Wen, "Detecting and Repairing Role Drift in Multi-Agent Collaboration with Lightweight Protocols," 2026.
49. Several entries are still not fully IEEE-complete because the original data does not include publisher, conference/journal name, page range, or publication status.
50. S. Kou, J. Jin, Z. Liu, et al., "Orthus: Autoregressive Interleaved Image-Text Generation with Modality-Specific Heads," arXiv preprint arXiv:2412.00127, 2024.
51. Y. Tuo, Y. Geng, and L. Bo, "AnyText2: Visual Text Generation and Editing with Customizable Attributes," arXiv preprint arXiv:2411.15245, 2024.
52. Y. Lee, J. Kim, J. Kim, et al., "CheckEval: A Reliable LLM-as-a-Judge Framework for Evaluating Text Generation Using Checklists," arXiv preprint arXiv:2403.18771, 2024.