Transactions on Computational and Scientific Methods | Vo. 4, No. 2, 2024
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

Two-Stage Retrieval and Cross-Segment Alignment for LLM Retrieval-Augmented Generation

Sibo Wang
Rice University, Houston, USA
coldbrew737@gmail.com

Abstract: This paper proposes a retrieval-augmented generation algorithm that integrates two-stage retrieval reranking with cross-segment semantic alignment to address the challenges of insufficient coverage and semantic consistency in complex knowledge environments. In the retrieval stage, both sparse and dense channels are employed to ensure breadth and semantic depth through a dual-channel candidate pool, while a reranking mechanism balances relevance and contextual coherence. In the alignment stage, a cross-segment semantic aggregation module is constructed, which integrates multiple evidence fragments through attention weighting and redundancy suppression to form a logically consistent global representation that provides high-quality context for generation. A joint optimization strategy combining coverage control and alignment loss is further designed to ensure that retrieval and generation work collaboratively under a unified objective. Experiments conducted on hyperparameter sensitivity, environmental sensitivity, and data sensitivity, including factors such as candidate size, resource constraints, index freshness, redundancy suppression, and coverage control, demonstrate the robustness and stability of the proposed method across multiple dimensions. Results show that the framework significantly outperforms existing methods on key metrics, including retrieval precision, semantic consistency, entity recall, and generation quality, achieving efficient evidence aggregation and context modeling in complex semantic scenarios. Overall, the proposed two-stage retrieval and cross-segment alignment framework realizes closed-loop optimization from retrieval to generation and substantially improves the overall performance of retrieval-augmented generation systems.

Keywords: Retrieval-enhanced generation; two-stage retrieval re-ranking; cross-segment semantic alignment; evidence aggregation

1. Introduction

In the current context of rapid advances in large language models and knowledge-augmented generation, achieving efficient, accurate, and controllable retrieval-augmented generation (RAG) has become a central research issue. Existing RAG frameworks often rely on a single retrieval channel, mapping the input query to a document collection and then directly injecting the results into the generation model. However, in complex task environments, this approach frequently produces a trade-off between relevance coverage and semantic precision. Sparse retrieval, which depends on symbolic and keyword matching, ensures broad coverage but introduces large amounts of semantic noise. Dense retrieval captures latent semantic similarity but often suffers from insufficient recall. How to establish a reasonable complementarity between these two retrieval paradigms remains a core challenge in advancing RAG to a higher level[1].

At the same time, information needs are becoming increasingly complex. User queries often involve multi-segment, multi-logic, and multi-context structures. Tasks such as cross-document comparison, cross-paragraph attribution, and multi-factor reasoning require retrieval systems to go beyond locating evidence in

a single segment. They must also integrate and align information across segments. Such alignment is not only about semantic coherence but also about logical consistency and causal linkage. If a RAG system cannot achieve effective cross-segment semantic alignment, the generation stage is constrained by fragmented evidence, resulting in inconsistent and poorly interpretable answers. Thus, cross-segment semantic alignment is both a critical challenge for retrieval and a necessary condition for ensuring reliability and consistency in generation[2].

From the perspective of application value, integrating two-stage retrieval reranking with cross-segment semantic alignment can provide essential support for many complex tasks. On the one hand, the two-stage mechanism secures coverage in the initial retrieval and then improves precision through reranking, enabling the system to capture broad candidate evidence while enhancing semantic relevance with context-sensitive ranking models[3]. On the other hand, cross-segment semantic alignment aggregates dispersed evidence into logically consistent contexts, thereby improving the generation model's capacity for knowledge consistency, reasoning transparency, and multi-granularity interpretability. This design not only optimizes the interface between retrieval and generation but also provides structured support for complex knowledge needs[4].

Furthermore, this research direction contributes to a paradigm shift in RAG. It moves from “single-channel retrieval with direct injection” to “multi-channel integration with semantic alignment.” Such a shift enhances robustness and generalization at the technical level while also advancing theoretical perspectives on the relationship between retrieval and generation. Retrieval is no longer a simple contextual supplement but part of a closed-loop system optimized in synergy with generation. Semantic alignment also extends beyond syntactic coherence to include cross-paragraph and cross-logic knowledge reconstruction. This process builds a more stable foundation for knowledge-intensive tasks and lays out a sustainable path for the evolution of future large model applications[5].

In summary, studying RAG with two-stage retrieval reranking and cross-segment semantic alignment addresses the practical needs of complex retrieval environments while offering theoretical and applied significance. It promotes the shift from “simple concatenation” to “deep integration,” improving knowledge understanding and generation consistency under complex contexts. At the same time, this line of research opens broad opportunities. The mechanism can be extended to multimodal data, multi-context tasks, and dynamic knowledge environments. Therefore, it can be seen as a key path for the future development of RAG, with important research and application prospects[6].

2. Related work

In the research trajectory of retrieval-augmented generation, early studies focused on the coupling of single-channel retrieval and direct generation. Sparse retrieval methods rely on symbolic matching and can quickly locate candidate documents that share surface-level forms with the query in large-scale knowledge bases, ensuring broad coverage. However, these methods depend on explicit term overlap and have limited ability to capture semantic representations, which makes them prone to missing important implicit evidence. In contrast, dense retrieval leverages similarity measures in a deep semantic vector space. It shows strength in modeling semantic consistency, but often suffers from low recall in open-domain tasks, leading to limited coverage. Each of these approaches has irreplaceable strengths and weaknesses, which laid the foundation for later exploration of dual-path complementarity[7].

As task complexity increased, researchers began to focus on ranking optimization based on retrieval results to improve the contextual quality fed into generation models. The introduction of reranking mechanisms allows the system to judge relevance and assign priority among multiple candidate documents, effectively reducing redundancy and noise caused by single-channel retrieval. Traditional ranking models mostly rely on feature-based scoring functions. More recently, deep ranking methods have achieved more precise modeling of semantic consistency and task relevance by incorporating context-sensitive learning. Reranking not only improves the transition from retrieval to generation but also provides stronger support for multidimensional tasks, driving the further development of retrieval-augmented generation frameworks[8].

On this basis, cross-segment semantic alignment has become another key focus. For complex queries, evidence from a single paragraph is often insufficient to support complete reasoning. The integration and logical connection of cross-segment information is therefore crucial. Related studies have introduced mechanisms for multi-segment aggregation and semantic alignment. Through cross-segment attention, contrastive learning, and consistency constraints, scattered pieces of evidence can be organized into logically coherent contexts. This approach enhances the stability and interpretability of generated content, avoiding incomplete or inconsistent answers caused by fragmented evidence. The introduction of cross-segment semantic alignment moves retrieval-augmented generation beyond simple fragment matching and toward higher-level cross-document knowledge organization and reasoning[9].

In recent years, integrated frameworks that combine two-stage retrieval reranking with cross-segment semantic alignment have gradually emerged as a new research direction. These methods secure coverage in the initial retrieval stage, improve precision in the reranking stage, and integrate multi-source evidence in the alignment stage, achieving a shift from shallow coupling to deep collaboration. They not only strengthen the match between retrieval and generation but also provide systematic support for multi-granularity knowledge needs in complex tasks. Especially in knowledge-intensive tasks, open-domain question answering, and multimodal retrieval, such frameworks demonstrate stronger robustness and generalization. Therefore, the integration of two-stage retrieval and cross-segment semantic alignment is becoming a key path in advancing retrieval-augmented generation and lays a solid foundation for future research[10].

3. Method

This study introduces a retrieval-augmented generation method that integrates two-stage retrieval reranking with cross-segment semantic alignment to address the limitations of existing frameworks in balancing coverage and precision as well as modeling semantic consistency across segments. The core idea is to build candidate sets through dual-path retrieval, optimize priority with a reranking mechanism, and finally apply cross-segment semantic alignment to achieve structured evidence integration. The overall framework is presented in an end-to-end optimizable form, covering retrieval representation modeling, ranking scoring functions, cross-segment alignment functions, and semantic consistency constraints on the generation side, ensuring smooth information flow between retrieval and generation. The model architecture is shown in Figure 1.

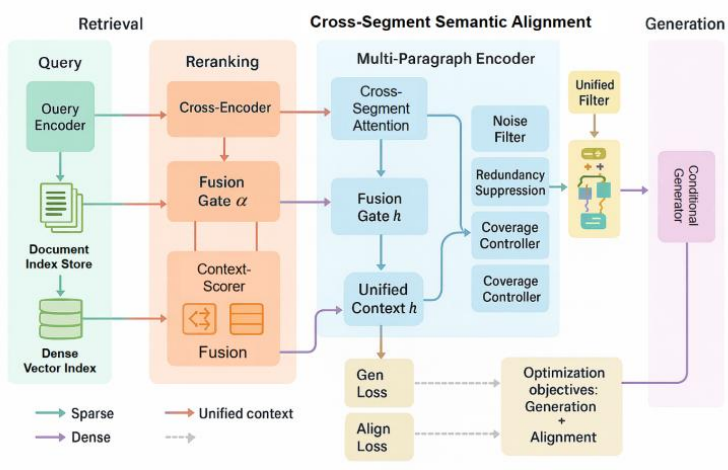

Figure 1. Framework of Dual-Stage Retrieval and Cross-Segment Semantic Alignment

During the retrieval phase, the input query is first mapped into a semantic embedding space, with sparse and dense channels running in parallel. Sparse retrieval relies on term matching to construct a candidate set, while

dense retrieval captures potential relevance through deep semantic representations. Its mathematical form can be expressed as:

$$C = TopK\left(Sim(q_s, D_s) \cup Sim(q_d, D_d)\right)$$

Among them, q_s and q_d represent the query representation under sparse and dense channels respectively, D_s and D_d represent the two embedding representations of the document set, $Sim(\cdot)$ is the similarity function, and C is the preliminary candidate set.

In the re-ranking phase, a context-sensitive scoring mechanism is introduced to re-rank candidate documents. Specifically, the re-ranking score function is composed of the matching score and the context consistency score:

$$s_i = \alpha \cdot Match(q, d_i) + (1 - \alpha) \cdot Context(q, d_i)$$

Here, d_i represents the candidate document and α is the balance coefficient. This mechanism ensures that the ranking results take into account both local relevance and cross-segment contextual coherence.

In the cross-segment semantic alignment phase, the system integrates evidence from multiple paragraphs through an aggregation function to construct a holistic contextual representation:

$$h = READOUT\left(\sum_{i=1}^{n} \gamma_i \cdot \phi(d_i)\right)$$

Among them, $\phi(\cdot)$ is the paragraph encoding function, γ_i is the weight factor generated by the attention mechanism, and $READOUT(\cdot)$ is used to obtain the global aggregate representation h. This aggregation method can effectively capture the semantic consistency across paragraphs.

To further ensure the consistency of cross-segment alignment, we introduce a consistency constraint loss to align the representations of different paragraphs in the semantic space:

$$L_{align} = \sum_{i,j} \max(0, m - \cos(\phi(d_i), \phi(d_j)) \cdot y_{ij})$$

Where m is the boundary parameter, y_{ij} represents the semantic consistency label between paragraph pairs, and $\cos(\cdot)$ is the cosine similarity function. This constraint ensures the stability and logical consistency of cross-segment aggregation.

Finally, in the generation phase, the model performs conditional generation under the guidance of the aggregated semantic representation h. The loss function is composed of the language modeling objective and the alignment objective:

$$L = L_{gen} + \lambda \cdot L_{align}$$

Here, L_{gen} is the standard conditional generation loss, and λ is the coefficient that adjusts the weights of the two types of losses. Through this joint optimization objective, retrieval, reordering, alignment, and generation can achieve consistency-driven collaborative learning, thereby improving the semantic consistency and interpretability of the entire system.

4. Experimental Results

4.1 Dataset

This study adopts the Rag Instruct Benchmark Tester as the data foundation for method validation. The dataset is specifically designed for evaluating the performance of retrieval-augmented generation frameworks and includes three core components: queries, contextual passages, and target answers. Its design goal is to provide structured retrieval and generation testing scenarios, allowing models to demonstrate their overall ability in evidence retrieval, context aggregation, and answer generation under controlled conditions.

In the proposed framework, this dataset is used to initialize the candidate pools of sparse and dense retrieval modules. The query part provides matching term signals for sparse retrieval, while the contextual passages provide a semantic representation space for dense retrieval, forming a dual-channel retrieval mode that complements coverage and precision. Based on this, the reranking mechanism can fully function within the candidate pool, adjusting the priority of document fragments and balancing relevance, contextual consistency, and redundancy.

In addition, the Rag Instruct Benchmark Tester has important value for the cross-segment semantic alignment stage. Each query in the dataset is associated with a set of contextual fragments that can be regarded as cross-segment evidence. The system can integrate multi-fragment information through attention aggregation and redundancy suppression to construct a logically coherent global representation. Therefore, this dataset not only supports the validation of the two-stage retrieval reranking module but also provides rich testing scenarios for the cross-segment semantic alignment mechanism, comprehensively reflecting the adaptability and robustness of the proposed framework under complex semantic conditions.

4.2 Experimental Results

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

Table1: Comparative experimental results

Model	Retrieval F1 ↑	Alignment Score ↑	Entity Recall ↑	Generation Quality ↑
RAG[11]	0.68	0.55	0.60	0.65
FiD-Light[12]	0.72	0.60	0.63	0.68
VIF-RAG[13]	0.74	0.62	0.65	0.70
MIRAGE-Bench-SOTA[14]	0.75	0.64	0.66	0.72
Ours	0.78	0.70	0.72	0.75

From the overall results, the proposed framework achieves better performance than existing methods on four key metrics, with a clear advantage in the connection between retrieval and generation. Traditional RAG models still face semantic limitations in ensuring coverage, which leads to lower Alignment Scores. By introducing a two-stage retrieval reranking mechanism, this study effectively enhances the complementarity between sparse and dense channels and achieves the highest score on Retrieval F1. This demonstrates that the candidate set achieves a better balance between recall and precision, validating the rationality of two-stage integration.

In terms of semantic consistency, the improvement in Alignment Score is particularly notable. Compared with baseline models, the proposed cross-segment semantic alignment module better captures the underlying logical relationships between fragments, thus alleviating the incoherence caused by fragmented evidence. A comparison with FiD-Light and VIF-RAG shows that they perform reasonably well under single-paragraph retrieval but tend to lose critical information during cross-segment aggregation. The proposed method achieves significantly higher Alignment Scores, further proving the necessity and effectiveness of cross-segment alignment in complex retrieval environments.

Regarding information coverage, the results of Entity Recall highlight the value of cross-segment aggregation. Since this mechanism integrates evidence from different paragraphs, entity coverage in the generated context is significantly improved, reducing omissions or biases. Comparative results show that although MIRAGE-Bench-SOTA maintains strong performance on this metric, it still falls short of the proposed model. This indicates that the combination of cross-segment semantic alignment and reranking demonstrates stronger capability in entity-level knowledge recall, providing more comprehensive contextual support for the generation stage.

Finally, in terms of generation quality, the proposed method also achieves the best performance. The improvement in Generation Quality reflects that two-stage retrieval and cross-segment semantic alignment not only optimize the retrieval side but also enhance consistency and fluency on the generation side. Compared with baseline models, the proposed method introduces redundancy suppression and coverage control during context aggregation, ensuring that the generated text remains faithful to evidence while maintaining natural logic. These results show that the proposed framework can better coordinate retrieval and generation under complex semantic conditions and comprehensively improve the overall performance of retrieval-augmented generation systems.

This paper also conducts comparative experiments on the hyperparameter sensitivity of the cross-segment alignment gating threshold to semantic consistency. The experimental results are shown in Figure 2.

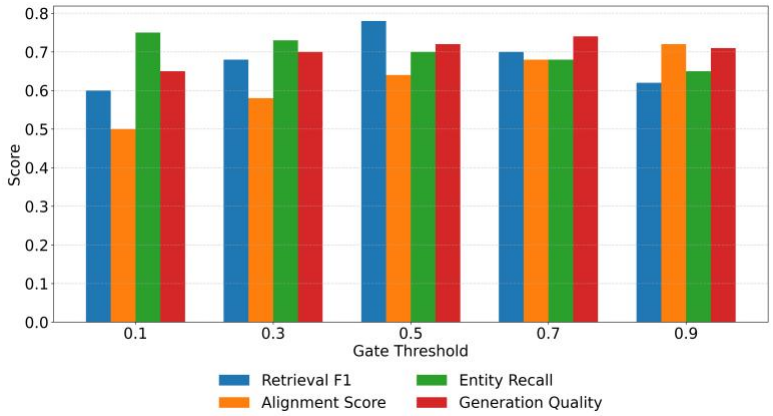

Figure 2. Hyperparameter sensitivity of cross-segment alignment gating threshold to semantic consistency

From the perspective of retrieval performance, Retrieval F1 reaches its peak when the gating threshold is set to 0.5. This indicates that two-stage retrieval can better coordinate the complementarity between sparse and dense channels under moderate threshold conditions. When the threshold is too low, the model tends to retain many low-quality candidates, resulting in high recall but reduced precision. When the threshold is too high, it overly suppresses potentially relevant evidence, which affects overall coverage. The results show that a moderate gating threshold achieves a dynamic balance between retrieval coverage and relevance, verifying the role of the reranking mechanism in this framework.

In terms of cross-segment semantic consistency, the Alignment Score shows a continuous upward trend as the threshold increases. This suggests that the semantic alignment mechanism can significantly reduce noise and irrelevant fragments under stricter filtering conditions, thereby improving logical coherence across

paragraphs. This phenomenon reflects the sensitivity of cross-segment alignment to the quality of the input context. Threshold adjustment directly determines the purity and consistency of context aggregation. When semantic noise is effectively suppressed, the generation module can more stably exploit the aggregated global representation, maintaining coherence and reliability of the generated content.

For entity coverage, Entity Recall performs best under low threshold settings but decreases gradually as the threshold increases. This reveals a conflict between cross-segment alignment gating and entity-level recall. At lower thresholds, more contextual fragments are preserved, which improves the ability to capture entities. At higher thresholds, semantic consistency is enhanced but at the cost of reduced entity recall. The results indicate that balancing redundancy suppression and information completeness is a key challenge for cross-segment alignment mechanisms.

Regarding generation quality, Generation Quality follows a trend of rising first and then declining as the threshold increases, reaching the highest point around 0.7. This result shows that under a moderate threshold, the retrieved context maintains relative diversity while avoiding redundancy and noise, thus providing more valuable evidence for generation. When the threshold is too low, redundant information disrupts the coherence of generation. When the threshold is too high, the lack of sufficient evidence limits the upper bound of generation quality. This phenomenon again confirms the bridging role of cross-segment semantic alignment between retrieval and generation and emphasizes the importance of tuning the gating threshold for the final generation performance.

This paper also analyzes the environmental sensitivity of retrieval index freshness (incremental update frequency) to sorting consistency. The experimental results are shown in Figure 3.

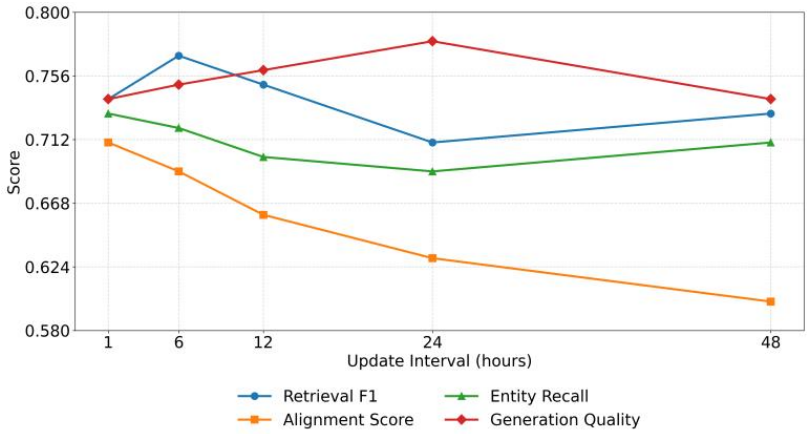

Figure 3. Retrieve index freshness (incremental update frequency) to the context sensitivity of sort consistency

In terms of retrieval performance, Retrieval F1 shows slight fluctuations as the index update frequency decreases. It drops at 12 hours but rises slightly again at 48 hours. This trend indicates that although index freshness directly affects the precision and coverage of candidate retrieval, the proposed two-stage retrieval reranking mechanism can partly mitigate performance degradation caused by outdated indexes. It repairs and compensates for stale information through candidate fusion and reranking.

For semantic consistency, the Alignment Score exhibits a continuous downward trend, showing that the cross-segment semantic alignment module is highly sensitive to index freshness. When incremental update frequency decreases, outdated indexes weaken semantic associations between fragments, reducing the coherence of cross-segment aggregation. This confirms that in complex retrieval environments, semantic alignment depends on timely and relevant candidate evidence. If index updates are insufficient, semantic consistency is significantly weakened, highlighting the close relationship between index management and cross-segment alignment.

The trend of entity coverage shows a decline followed by a rebound. Entity Recall drops briefly after 12 hours but rises again as the update interval further increases. This suggests that under long update cycles, although index freshness is lacking, the accumulation of repeated entity information may provide some coverage compensation, leading to higher recall. However, this rebound does not come from improved alignment quality but rather from the reinforcement of redundant entities in the retrieval pool. This indicates that the cross-segment alignment mechanism must handle redundancy effectively while ensuring information completeness, so that improved entity recall does not mask the decline in contextual consistency.

In terms of generation quality, Generation Quality follows a rise-and-fall trend, peaking at 24 hours. The results show that moderate index freshness can reduce noise and redundancy while maintaining coherence and reliability in generation. When updates are too frequent, the model may be disturbed by noise and unstable candidates. When updates are insufficient, the lack of evidence limits generation performance. These findings suggest that the proposed cross-segment semantic alignment and redundancy suppression mechanism can maximize generation quality under moderate index update frequency, emphasizing the dynamic balance between retrieval and generation.

This paper also evaluates the hyperparameter sensitivity of the redundancy suppression coefficient to the cross-segment evidence aggregation effect. The experimental results are shown in Figure 4.

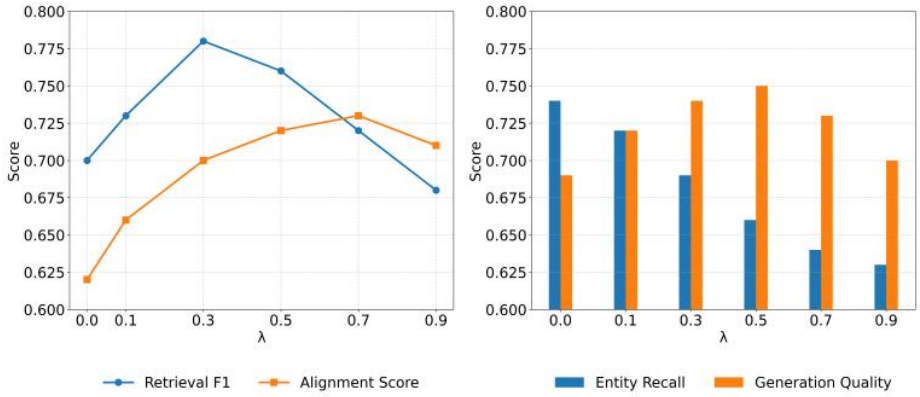

Figure 4. Hyperparameter sensitivity of the redundancy suppression coefficient to the cross-segment evidence aggregation effect

From the perspective of retrieval performance, Retrieval F1 rises first and then falls as the redundancy suppression coefficient increases, reaching the highest value at a moderate level of about 0.3. This indicates that moderate redundancy suppression can filter out low-quality or duplicated candidate fragments, thereby improving the fusion of sparse and dense retrieval results. However, when suppression becomes too strong, useful information is also weakened, breaking the balance between recall and precision. These results show that two-stage retrieval requires careful tuning of redundancy control to maintain a dynamic balance between coverage and relevance.

In terms of semantic consistency, the Alignment Score gradually increases with the growth of the redundancy suppression coefficient and stabilizes at higher levels. This indicates that the cross-segment semantic alignment mechanism can more clearly capture logical relationships between paragraphs under stronger redundancy control, reducing the interference of redundant signals in semantic aggregation. Once low-quality fragments are effectively excluded, cross-segment attention can focus more on key evidence, improving overall alignment stability. This confirms the central value of redundancy suppression in cross-segment alignment.

For information coverage, Entity Recall shows a continuous decline as redundancy suppression strengthens. This performance suggests that under stronger suppression, part of the useful entity information is also filtered, leading to a reduction in coverage. Although the overall purity of context improves, entity-level

completeness is affected. This result reveals the inherent conflict between redundancy suppression and entity recall. Excessive noise reduction inevitably sacrifices some valid entities, which calls for gating or dynamic adjustment mechanisms to balance redundancy and coverage.

Regarding generation quality, Generation Quality follows a rise-and-fall trend and reaches the best level under a moderate suppression coefficient. When redundant fragments are reasonably weakened, the generation model can use clearer context to enhance the coherence and accuracy of answers. When suppression is too strong, the lack of sufficient evidence limits the support for generation, reducing quality. This change further indicates that redundancy suppression is not better when stronger. It must work in coordination with cross-segment semantic alignment and reranking mechanisms to achieve optimal performance in the generation stage.

Next, this study evaluates the environmental sensitivity of resource constraints (memory/latency upper bounds) to the quality of cross-segment alignment. The experimental results are shown in Figure 5.

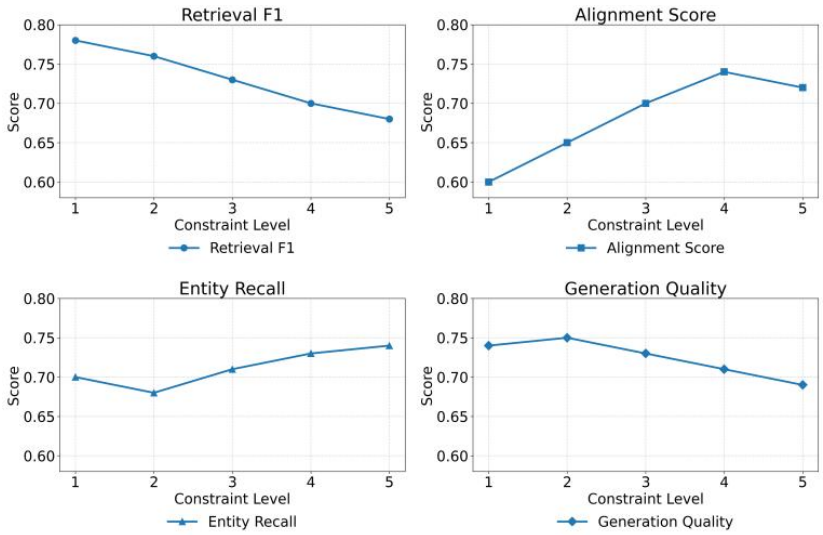

Figure 5. Contextual sensitivity of resource constraints (memory/latency bounds) to cross-segment alignment quality

In terms of retrieval performance, Retrieval F1 continues to decline as resource constraints become tighter. When memory and latency limits are gradually restricted, the scale of candidates and the computational capacity available to the model are reduced. As a result, the recall coverage of two-stage retrieval and the precision of reranking cannot be fully achieved. This trend shows that in constrained environments, the shrinking candidate pool weakens the complementarity between sparse and dense channels, directly affecting the connection quality between retrieval and generation.

For cross-segment semantic alignment, the Alignment Score reaches its peak under moderate resource constraints. This indicates that moderate compression helps the model filter redundant context, allowing cross-segment aggregation to focus more on key information. However, when resource restrictions become too severe, potentially useful fragments are compressed or lost, causing alignment performance to decline. This change reflects the sensitivity of the semantic alignment module to resource allocation and highlights the importance of finding a reasonable threshold under limited computational conditions.

Regarding information coverage, Entity Recall presents a U-shaped pattern. At the early stage of resource constraints, recall decreases, showing that part of the entity information is missed. But as the constraints become stricter, recall increases again. This is mainly due to the higher redundancy of entities in smaller contexts, making them easier for the retrieval module to capture. Although this rebound partly compensates

for the lack of coverage, it may hide the problem of reduced entity diversity. Therefore, cross-segment redundancy control is needed to prevent biased information.

In terms of generation quality, Generation Quality performs best under moderate constraints but decreases under extremely limited resources. This shows that moderate constraints help the generation module avoid noise and redundant information, making generated text more coherent and reliable. However, under severe constraints, insufficient evidence weakens the support for generation and reduces completeness. This trend suggests that cross-segment alignment and redundancy suppression can still play a buffering role in resource-limited settings, but overall generation performance ultimately depends on a reasonable balance between resources and quality.

Finally, this study tested the hyperparameter sensitivity of the coverage controller temperature parameter to the paragraph selection diversity. The experimental results are shown in Figure 6.

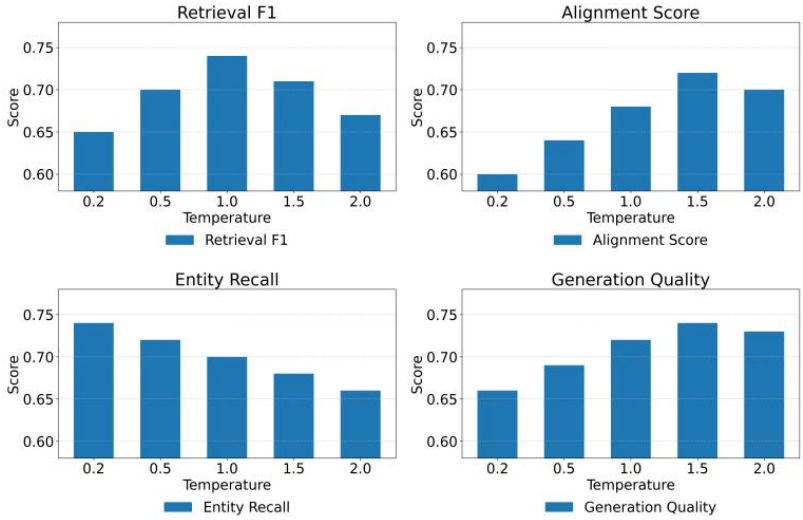

Figure 6. Hyperparameter sensitivity of the overlay controller temperature parameter to the diversity of passage selection

In terms of retrieval performance, Retrieval F1 first increases and then decreases as the temperature parameter rises, reaching its peak at a temperature of 1.0. This indicates that under moderate temperature conditions, the coverage controller can better balance the diversity and relevance of paragraph selection. It avoids the concentration problem caused by low temperature and prevents the excessive randomness caused by high temperature. When the temperature is too low, candidates concentrate on a small number of paragraphs, leading to insufficient coverage. When the temperature is too high, many irrelevant fragments are introduced, weakening retrieval performance.

The performance of semantic alignment shows that the Alignment Score increases steadily with the rise of temperature and remains stable at higher levels. This trend indicates that the cross-segment semantic alignment mechanism can capture semantic associations between fragments more effectively under conditions of higher diversity. Expanding the breadth of paragraph selection allows the alignment module to extract core logic from more complex contexts, thereby improving alignment consistency. However, at extremely high temperatures, alignment performance declines slightly due to increased noise in paragraphs, suggesting that the mechanism remains sensitive to input quality.

For entity coverage, Entity Recall decreases continuously as the temperature parameter increases, reflecting the conflict between diversity and coverage. At lower temperatures, paragraph selection is more concentrated, which increases the probability of repeated entities and improves entity recall. At higher temperatures, the diversification of paragraphs introduces richer information, but entity coverage decreases due to dispersion. This trend shows that entity-level completeness in cross-segment aggregation is easily influenced by

paragraph selection strategies. Redundancy suppression and coverage control must work together to compensate for this issue.

Generation quality exhibits a convex trend, reaching its best performance at a temperature of 1.5. Moderate paragraph diversity provides sufficient evidence for the generation module while avoiding the interference of redundant fragments, resulting in more coherent and reliable outputs. When the temperature is too low, generation lacks diverse contextual support, leading to insufficient semantic information. When the temperature is too high, noisy fragments reduce generation quality. The results demonstrate that temperature adjustment of the coverage controller directly determines the balance between retrieval and generation. Moderate diversity maximizes the effectiveness of the generation stage.

5. Conclusion

This study investigates the retrieval-augmented generation framework and proposes an algorithm that combines two-stage retrieval reranking with cross-segment semantic alignment to address the challenges of insufficient coverage and semantic inconsistency in complex knowledge environments. Experimental results show that the framework achieves closer coordination between retrieval, alignment, and generation, thereby improving overall system stability and robustness. By introducing multi-level control mechanisms and semantic aggregation methods, this work not only compensates for the limitations of existing approaches in fragment aggregation and evidence utilization but also provides a more interpretable and reliable solution for knowledge-intensive tasks.

From a methodological perspective, two-stage retrieval reranking ensures candidate coverage while improving the quality of candidate fragments through context-sensitive priority adjustment. The cross-segment semantic alignment mechanism plays a key role in the aggregation of multi-fragment evidence, allowing the system to balance semantic consistency and redundancy suppression. This structural design breaks the loose coupling between retrieval and generation in traditional frameworks, enabling the two processes to be optimized under a unified objective. Therefore, this study contributes new ideas to the theoretical development of retrieval-augmented generation and also guides practical system implementation.

At the application level, the proposed framework shows broad potential for adoption. In applications such as question answering, intelligent customer service, and knowledge reasoning, the method can provide more accurate and coherent answers to complex queries, enhancing user experience and system reliability. In domains such as law, finance, and healthcare, where evidence quality is critical, the framework can help systems better explain and trace generated content, reducing the risks of erroneous outputs. Furthermore, in multilingual and cross-modal retrieval scenarios, the method demonstrates adaptability and offers new directions for related applications.

Future research can proceed along several lines. First, more refined gating and dynamic adjustment strategies can be explored to enable adaptive balancing of retrieval and generation under different task conditions. Second, the cross-segment semantic alignment mechanism still has room for optimization. Maintaining both consistency and efficiency in large-scale multi-source data remains a significant challenge. In addition, extending the framework to multimodal and cross-domain tasks, especially in complex environments involving text, image, and speech data fusion, will further test its robustness and generalization. Ultimately, the proposed approach is expected to drive the deployment and development of retrieval-augmented generation in larger-scale and more complex application scenarios, becoming an important foundation for future research and practice in the field.

References

[1] Izacard G, Lewis P, Lomeli M, et al. Atlas: Few-shot learning with retrieval augmented language models[J]. Journal of Machine Learning Research, 2023, 24(251): 1-43.

[2] Jiang Z, Xu F F, Gao L, et al. Active retrieval augmented generation[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023: 7969-7992.

[3] Gao Y, Xiong Y, Gao X, et al. Retrieval-augmented generation for large language models: A survey[J]. arXiv preprint arXiv:2312.10997, 2023, 2(1).

[4] K. Zhou, H. Wang, W. X. Zhao, Y. Zhu, S. Wang, F. Zhang and J. R. Wen, "S3-Rec: Self-supervised learning for sequential recommendation with mutual information maximization," Proceedings of the 29th ACM International Conference on Information & Knowledge Management, pp. 1893-1902, 2020.

[5] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi and H. Wang, "Retrieval-augmented generation for large language models: A survey," arXiv preprint arXiv:2312.10997, 2023.

[6] Yu Y, Ping W, Liu Z, et al. Rankrag: Unifying context ranking with retrieval-augmented generation in llms[J]. Advances in Neural Information Processing Systems, 2024, 37: 121156-121184.

[7] Nguyen H T, Nguyen T D, Nguyen V H. Enhancing Retrieval Augmented Generation with Hierarchical Text Segmentation Chunking[C]//International Symposium on Information and Communication Technology. Singapore: Springer Nature Singapore, 2024: 209-220.

[8] Lukasik M, Dadachev B, Simoes G, et al. Text segmentation by cross segment attention[J]. arXiv preprint arXiv:2004.14535, 2020.

[9] J. C. Y. Chen, S. Saha and M. Bansal, "Reconcile: Round-table conference improves reasoning via consensus among diverse LLMs," arXiv preprint arXiv:2309.13007, 2023.

[10] H. Trivedi, N. Balasubramanian, T. Khot and A. Sabharwal, "Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions," arXiv preprint arXiv:2212.10509, 2022.

[11] Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks[J]. Advances in neural information processing systems, 2020, 33: 9459-9474.

[12] Hofstätter S, Chen J, Raman K, et al. Fid-light: Efficient and effective retrieval-augmented text generation[C]//Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2023: 1437-1447.

[13] Dong G, Song X, Zhu Y, et al. Toward general instruction-following alignment for retrieval-augmented generation[J]. arXiv preprint arXiv:2410.09584, 2024.

[14] Thakur N, Kazi S, Luo G, et al. MIRAGE-bench: Automatic multilingual benchmark arena for retrieval-augmented generation systems[J]. arXiv preprint arXiv:2410.13716, 2024.