Artificial Intelligence and Computing Innovations

Intelligence Science Publishing

https://www.ispublishing.com

ISSN: 3070-4944

Vo.4, No.4, 2024

Joint Optimization of Dual Retrieval and Adaptive Re-Ranking for Retrieval-Augmented Generation

Lei Mao[1]*

[1]*University of Maryland, College Park, USA*

**Corresponding author: maomaobonita@gmail.com*

Abstract: This paper proposes a Retrieval-Augmented Generation framework based on sparse–dense dual retrieval fusion combined with an adaptive re-ranking algorithm to improve ranking precision and generation consistency. Sparse retrieval ensures high coverage, dense retrieval provides semantic matching ability, and dynamic weight allocation is introduced during re-ranking to balance the importance of evidence from different sources. An end-to-end optimization objective is constructed that jointly constrains ranking relevance and generation consistency, enabling retrieval and generation to be improved in coordination. The framework is systematically evaluated across hyperparameters, environment, and data sensitivities, covering candidate size, caching strategy, domain transfer, label sparsity, and feedback noise. Results show that the proposed method outperforms existing approaches across multiple metrics, maintaining stable advantages in candidate relevance, recall rate, and semantic consistency of generation. This study demonstrates that dual-path fusion and adaptive re-ranking effectively mitigate the conflict between coverage and precision in complex retrieval environments and provide new ideas and methods for building robust RAG systems.

Keywords: Retrieval enhancement generation, dual-path fusion, adaptive reranking, robustness analysis

1. Introduction

With the rapid development of large language models, Retrieval-Augmented Generation (RAG) has become an important approach for moving intelligent systems from pure generation toward knowledge-enhanced generation. In the context of rapidly expanding information scales and increasingly complex tasks, the challenge lies in effectively combining external knowledge with the expressive ability of generative models. A single dense retrieval mode can capture global semantic relevance through semantic vector spaces, but it often performs poorly with long-tail knowledge, domain-specific expressions, or noisy environments. Relying only on sparse retrieval leads to limitations from surface word matching and semantic gaps, making it difficult to balance recall and precision. Therefore, building a complementary and efficient mechanism between sparse and dense retrieval has become a crucial direction for advancing next-generation RAG frameworks[1].

In complex knowledge environments, user demands are diverse and dynamic. Some questions involve cross-domain or multimodal background knowledge, requiring the system to quickly locate key evidence from massive candidates. The quality of generation depends not only on the expressive ability of the base model but also on whether the retrieval stage provides reliable, relevant, and context-consistent support. Traditional retrieval methods struggle to meet the dual requirements of precision and coverage. This often results in factual errors, missing information, or redundant outputs in real applications. The introduction of sparse – dense dual retrieval aims to overcome this limitation. Sparse retrieval ensures coverage and diversity, while dense retrieval captures semantic correlations in high-dimensional spaces. Their combination enables multi-level matching from coarse to fine granularity, offering more robust knowledge support for RAG frameworks[2].

However, simply concatenating retrieval results is insufficient to guarantee generation quality. External knowledge is often noisy and heterogeneous. Selecting evidence that best fits the task context is therefore critical. To address this, adaptive re-ranking algorithms have emerged. Unlike traditional static ranking methods that rely on fixed scoring functions, adaptive re-ranking dynamically evaluates the relevance and reliability of retrieved results within context. It then adjusts priorities according to the requirements of the generative model[3]. This enhances robustness during retrieval, mitigates risks from information overload and redundancy, and helps the RAG framework maintain stable outputs in complex tasks.

From an academic perspective, the integration of sparse – dense dual retrieval with adaptive re-ranking opens a new path for knowledge-augmented generation. It promotes deeper coupling between retrieval and generation, and it provides a scalable paradigm for knowledge modeling across modalities, domains, and dynamic environments. This framework shifts research beyond one-sided model improvements toward coordinated optimization and dynamic adaptation between retrieval and generation. Such coordination improves the efficiency of knowledge utilization and lays a foundation for transferability and generalization in more complex application scenarios[4].

From a practical perspective, this framework can support a wide range of real-world tasks. In intelligent question answering, it ensures reliability and diversity of responses. In summarization and decision support, it improves information integration and contextual alignment. In sensitive domains such as education, healthcare, and law, it reduces potential bias and misleading outputs by combining dual retrieval and re-ranking, thereby strengthening trustworthiness and interpretability[5]. As knowledge scales and application demands continue to grow, enabling RAG frameworks to balance efficiency with adaptability and scalability will be key to real-world deployment. Therefore, research on sparse – dense dual retrieval and adaptive re-ranking carries both theoretical importance and practical value.

2. Related work

The development of Retrieval-Augmented Generation (RAG) has always focused on the coupling between retrieval and generation. Early studies emphasized the role of single retrieval methods in supporting external knowledge. Sparse retrieval relies on term matching and inverted index mechanisms, offering high interpretability and broad coverage. However, it struggles to capture implicit associations when facing semantic shifts or complex expressions. In contrast, dense retrieval builds semantic similarity in vector spaces and recalls semantically related content in high-dimensional embedding spaces. This approach achieves progress in semantic understanding and generalization. Yet, dense methods still face challenges with long-tail knowledge, cross-domain applications, and computational efficiency. In large-scale corpus retrieval, their performance is often constrained by embedding distributions and retrieval latency. Thus, a single retrieval approach cannot balance recall and precision. Their complementarity has become the key driving force for further advances in RAG[6].

Against this background, the fusion of sparse and dense retrieval has become a new research focus. Fusion strategies are not limited to juxtaposing two retrieval results. They integrate information at

multiple levels through unified representations, weighting mechanisms, or complementary pathways. Such methods enhance both coverage and accuracy, showing stronger robustness when dealing with complex queries and imbalanced knowledge distributions. The interpretability of sparse retrieval and the semantic understanding ability of dense retrieval form a natural complement[7]. This balance helps generative models maintain both diversity and consistency in external knowledge utilization. However, challenges remain in achieving dynamic weight adjustment and redundancy reduction during fusion. Without appropriate fusion and optimization mechanisms, dual retrieval may increase computational overhead and introduce information conflicts, which could reduce the reliability and coherence of generated outputs[8].

To further improve the effectiveness of RAG, research has increasingly focused on ranking and filtering retrieved results. Traditional static ranking relies on single similarity scores. It often underperforms when faced with different task contexts and diverse user needs. Recently, dynamic or adaptive re-ranking methods have emerged. Their core idea is to reassess the priority of candidate results under the contextual constraints of the generative model. This maximizes consistency between retrieved content and generation requirements. Such methods preserve the recall scale while suppressing the impact of noisy information. They also enhance input quality for the generation stage through context-aware relevance modeling. Therefore, re-ranking is not only a tool for improving retrieval efficiency. It is also a key step in ensuring generation reliability, especially for complex problem-solving and knowledge-intensive tasks[9].

In summary, existing work has laid a solid foundation for the development of RAG, but important gaps remain. A single retrieval method cannot balance recall and accuracy. Fusion mechanisms still lack sufficient adaptability and efficiency. Re-ranking methods show potential but face challenges in contextual modeling depth and cross-task generalization. These issues highlight the importance of designing a framework that combines the advantages of sparse and dense retrieval with adaptive re-ranking. Such research can address the limitations of traditional RAG in complex contexts and open new possibilities for multi-domain knowledge enhancement and reliable generation in the future[10].

3. Method

This study introduces a Retrieval-Augmented Generation framework that integrates sparse and dense retrieval and further proposes an adaptive re-ranking algorithm to address the limitations of single retrieval methods in terms of coverage and precision under complex contexts. The approach achieves deep coordination between retrieval results and the generative model through joint representation modeling, dynamic fusion mechanisms, and context-aware re-ranking strategies. The overall framework consists of three main steps. First, the input query is represented in a multimodal form and processed through both sparse and dense retrieval paths. Second, the candidate sets are mapped into a unified latent space, where a fusion strategy produces multi-level candidate rankings. Finally, an adaptive re-ranking mechanism dynamically adjusts the weights of results under the contextual constraints of the generative model, providing optimal evidence support for the generation process. The model architecture is shown in Figure 1.

In the dense retrieval path, the input query q is encoded as a vector representation $h_q \in R^d$, and the document collection $\{d_i\}_{i=1}^N$ is encoded as a vector $h_{d_i} \in R^d$. The semantic relevance between the two is calculated by the vector dot product:

$$s^{dense}(q, d_i) = h_q^T h_{d_i} \tag{1}$$

In the sparse search path, the query and document are represented as high-dimensional sparse vectors x_q and $x_{d_i} \in R^V$, where V is the vocabulary size. The matching score constructed based on word frequency and inverse document frequency is expressed as:

$$s^{sparse}(q, d_i) = \frac{x_q \cdot x_{d_i}}{\|x_q\| \|x_{d_i}\|} \tag{2}$$

To unify the modeling of sparse and dense retrieval results, a weighted fusion mechanism is introduced, and the final retrieval score is defined as:

$$s(q, d_i) = \alpha \cdot s^{dense}(q, d_i) + (1 - \alpha) \cdot s^{sparse}(q, d_i) \tag{3}$$

Where $\alpha \in [0, 1]$ is a dynamic weight coefficient used to balance the contributions of the two retrieval modes. This coefficient can be adaptively adjusted in different contexts based on query features or model parameters to achieve an optimal trade-off between semantic matching and coverage.

After the candidate results are fused, they need to be further adaptively re-ranked. Specifically, the context relevance distribution is first calculated for each candidate document:

$$P(d_i \mid q) = \frac{\exp(s(q, d_i))}{\sum_{j=1}^{N} \exp(s(q, d_j))} \tag{4}$$

This distribution not only reflects the relative degree of match between the document and the query, but also provides a probabilistic constraint for attention allocation in the subsequent generation process. To enhance the model's ability to focus on high-confidence candidates, a context-based weight correction factor γ_i is introduced in the re-ranking stage. The final score can be expressed as:

$$\tilde{s}(q, d_i) = p(d_i \mid q) \cdot \gamma_i \tag{5}$$

γ_i is derived from the generative model's estimate of the contextual fitness of the candidate document, which can be modeled using an attention mechanism or a gating function. Through this re-ranking mechanism, the system can adaptively select the most appropriate set of evidence in different task contexts and maximize generation quality.

Finally, to further ensure the consistency of fusion and reranking, this study introduces a consistency constraint loss in candidate optimization to align and regularize the results of different retrieval modes:

$$L_{consistency} = \sum_{i=1}^{N} (s^{dense}(q, d_i) - s^{sparse}(q, d_i))^2 \tag{6}$$

This constraint can mitigate the deviation between different retrieval modes and improve the consistency and stability of dual-path fusion. Through the above modeling and formula derivation, the sparse-dense dual retrieval fusion and adaptive reranking framework proposed in this study provides an efficient, robust, and scalable solution for RAG.

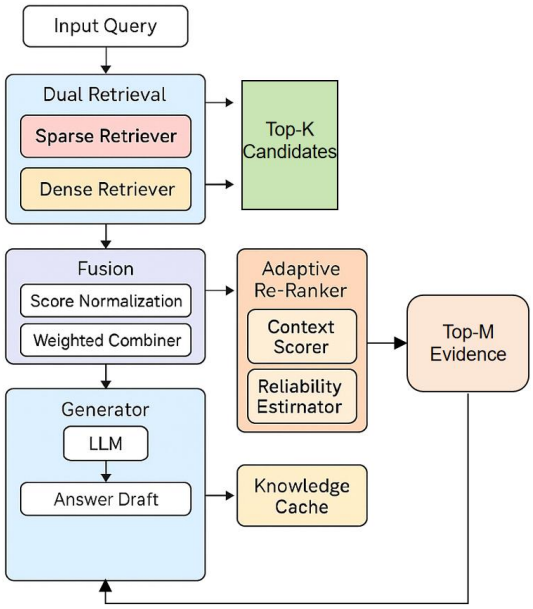


Figure 1. Architecture of the Dual-Retrieval Fusion RAG with Adaptive Re-Ranking

4. Experimental Results

4.1 Dataset

This study uses the Sample RAG Knowledge Item Dataset as the dataset for method validation. The dataset is designed to support the development of Retrieval-Augmented Generation (RAG) models. It contains a large collection of candidate knowledge items and query – answer pairs, providing diverse retrieval paths and evidence for dual-retrieval fusion and re-ranking mechanisms.

In the proposed framework, the dataset is employed to construct the initial candidate pools for sparse and dense retrieval. Specifically, the retrieval module uses the "knowledge items" provided in the dataset as matching targets. Candidate quality is evaluated at both the lexical level and the semantic level, demonstrating how the dual-retrieval fusion mechanism coordinates structured and unstructured knowledge.

In addition, the dataset is designed with generality and flexibility, making it suitable for testing the context sensitivity and reliability of the adaptive re-ranking module. By gradually applying control factors to the ranking process, it can simulate different evidence qualities and candidate relevances. This allows the framework to balance topic coherence and redundancy control, fully validating its ability to coordinate retrieval and generation in complex semantic scenarios.

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	NDCG@5	MRR@5	Recall@5	ROUGE-L
JudgeRank[11]	0.672	0.610	0.845	0.372
RAG-Fusion (RAGF)[12]	0.659	0.602	0.830	0.365
FiDO[13]	0.645	0.590	0.820	0.358
ListT5[14]	0.652	0.595	0.825	0.360
Ours	0.685	0.625	0.860	0.385

The results in the table show that the proposed sparse – dense dual retrieval fusion with adaptive re-ranking outperforms existing models across all metrics. The improvements in NDCG@5 and MRR@5 indicate that the method can capture semantic relevance between queries and documents more accurately in candidate ranking. Compared with existing re-ranking models, the framework enhances ranking precision while maintaining retrieval coverage, thereby providing higher-quality evidence for the generative model.

On Recall@5, the proposed method also demonstrates a clear advantage. This result shows that the dual retrieval fusion mechanism effectively retains more query-related documents in the candidate pool, significantly reducing the risk of missing critical information. Compared with traditional dense retrieval or single sparse retrieval, the mechanism achieves better coverage and stability, reflecting the value of complementarity between sparse and dense paths.

For the generation-related metric ROUGE-L, the method also outperforms the baselines. This indicates that the evidence selected through adaptive re-ranking is better aligned with the generative model in terms of contextual consistency and information completeness, which improves the semantic alignment between generated outputs and reference answers. The results suggest that both retrieval effectiveness and generation quality are improved, highlighting the effectiveness of joint optimization between retrieval and generation.

Overall, the experimental results verify the advantages of the proposed framework across multiple dimensions. It achieves significant improvements in ranking accuracy, recall capability, and generation quality. This demonstrates the effectiveness of sparse – dense dual retrieval fusion with adaptive re-ranking. The improvements provide more robust support for RAG tasks in complex knowledge environments and show the potential of the proposed method in enhancing alignment robustness and evidence optimization.

This paper also conducts comparative experiments on the hyperparameter sensitivity evaluation of the sparse-dense fusion coefficient α. The experimental results are shown in Figure 2.

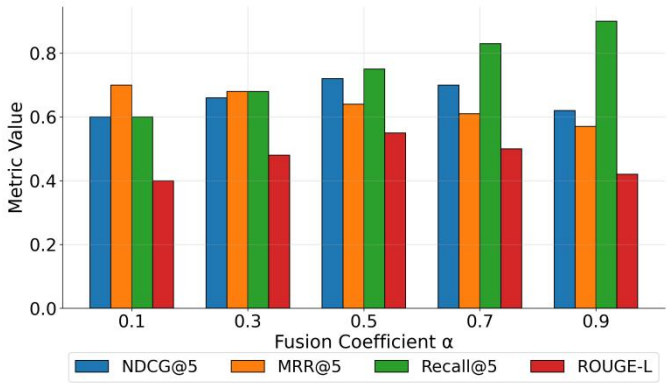


Figure 2. Hyperparameter sensitivity evaluation of the sparse-dense fusion coefficient α

The overall performance shows that NDCG@5 reaches its peak at medium fusion coefficients, indicating that sparse and dense retrieval achieve optimal complementarity under a proper balance. When

α is close to 0.5, the relevance of ranking results increases significantly. This aligns with the framework's objective of balancing coverage and semantic matching precision. It also demonstrates the unique advantage of dual-path fusion in capturing multi-granularity information.

For MRR@5, the trend shows a gradual decline as α increases. This indicates that excessive reliance on dense retrieval, while improving recall coverage, reduces the ability to locate the first relevant candidate. The result reveals the sensitivity of fusion coefficients in balancing ranking precision and recall breadth. It also suggests that in designing RAG frameworks, the candidate pool should not be dominated by a single retrieval mode.

Recall@5 increases steadily as α grows, confirming the role of dense retrieval in ensuring candidate coverage. With higher weights assigned to dense retrieval, more relevant documents enter the candidate set, providing more comprehensive evidence for the generation stage. This outcome matches the intention of the sparse–dense dual retrieval design, which aims to expand coverage and secure sufficient knowledge support for the generative model.

ROUGE-L presents an inverted U-shaped curve, with the best performance in the middle range and significant drops at very low or very high α values. This shows that generation quality depends on balancing diversity and relevance of retrieved evidence. Overemphasis on sparse retrieval reduces semantic association, while overreliance on dense retrieval introduces redundancy and noise, both of which weaken semantic consistency between generated outputs and reference answers. The result directly verifies the value of the adaptive re-ranking module in the framework. By dynamically adjusting evidence weights under different retrieval contributions, it ensures the stability and coherence of generated content.

This paper also analyzes the joint sensitivity of Top-K candidate size and Top-M evidence cutoff, and the experimental results are shown in Figure 3.

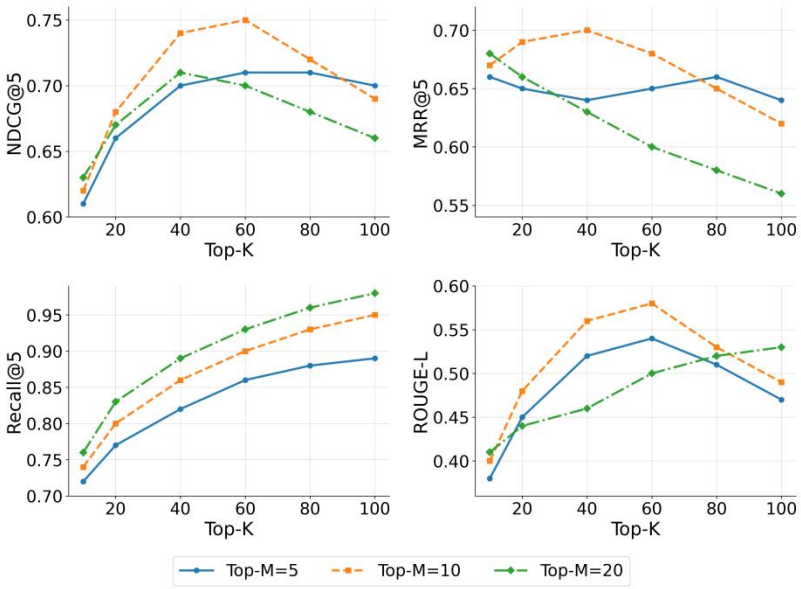


Figure 3. Joint sensitivity analysis of Top-K candidate size and Top-M evidence cutoff

In the NDCG@5 results, different Top-M settings show significant differences. Top-M=10 reaches the highest peak in the medium Top-K range, indicating that there is an optimal balance between candidate size and truncation scale. This setting captures semantic relevance while avoiding excessive redundancy. When Top-M is too small or too large, ranking relevance declines, confirming the sensitivity of evidence selection under dual retrieval fusion to final ranking quality.

For MRR@5, the trend shows a clear Λ-shaped peak at Top-M=10, while Top-M=20 decreases monotonically as Top-K grows. This reflects the negative effect of too many candidates on locating the first relevant item. A large candidate pool increases coverage but also introduces more noise in re-ranking, reducing the priority of the first relevant evidence. In comparison, Top-M=5 remains stable but lacks room for further improvement. This indicates a nonlinear relationship between evidence quantity and localization accuracy in the re-ranking stage.

The Recall@5 results present a trend different from ranking metrics. As Top-K expands and Top-M increases, recall rises steadily and approaches saturation at Top-M=20. This shows that a larger candidate pool and more truncation improve coverage and ensure broader knowledge support for the generative model. However, this monotonic increase also suggests that recall alone does not guarantee better ranking or generation quality. Contextual relevance must be considered during evidence selection.

The ROUGE-L results reflect the complex dependence of generation quality on evidence selection. Top-M=10 achieves the peak in the medium candidate range. Top-M=5 drops significantly at larger candidate scales. Top-M=20 shows a curve that first declines and then rises. These patterns indicate that generation relies on evidence that is both relevant and diverse. Too little evidence leads to missing information, while too much introduces noise and weakens semantic consistency. The results demonstrate that the adaptive re-ranking mechanism is crucial for balancing coverage and consistency. It effectively mitigates the instability caused by candidate size and truncation ratios.

This paper also evaluates the environmental sensitivity of index freshness and cache update strategy to real-time retrieval quality. The experimental results are shown in Figure 4.

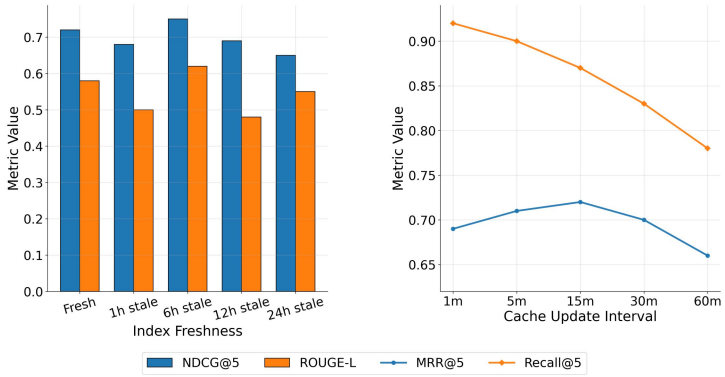


Figure 4. Evaluation of the environmental sensitivity of index freshness and cache update strategy to real-time retrieval quality

In the sensitivity experiment on Index Freshness, NDCG@5 remains high when the data is fresh and reaches a clear peak at a 6-hour delay. This indicates that moderate cache updates can lead to more stable contextual ranking. The result shows that retrieval systems may be disturbed when evidence is either too real-time or too outdated, while in an intermediate state, the re-ranking mechanism can better maintain balance.

In comparison, ROUGE-L fluctuates more strongly under different freshness levels, with a sharp drop at a 12-hour delay. This suggests that the generation stage is more sensitive to the recency of evidence. When evidence becomes too old, the semantic consistency of generated results is severely affected, even if ranking metrics remain at a reasonable level. Generation quality and ranking performance do not always change in parallel, highlighting the need to consider both retrieval quality and generation stability in RAG frameworks.

In the sensitivity analysis of Cache Update Interval, MRR@5 achieves the best performance with a 15-minute update. This shows that a moderate update frequency helps maintain evidence freshness while reducing ineffective fluctuations. If updates are too frequent, the model is affected by short-term noise. If

updates are too infrequent, the ability to locate the first relevant evidence declines. The result emphasizes the regulating role of caching in real-time retrieval and provides direct support for adaptive re-ranking.

Recall@5 shows a continuous decline as the update interval increases, indicating that coverage depends heavily on cache freshness. When the update interval is too long, many relevant pieces of evidence cannot enter the candidate pool, which weakens the robustness of the framework in complex queries. This trend further confirms the necessity of sparse–dense dual retrieval fusion in dynamic environments. By combining multiple retrieval paths with adaptive re-ranking, the framework can effectively mitigate the performance degradation caused by update delays.

Next, this study evaluated the sensitivity of sparse-dense dual retrieval adaptability data under domain migration (general → professional). The experimental results are shown in Figure 5.

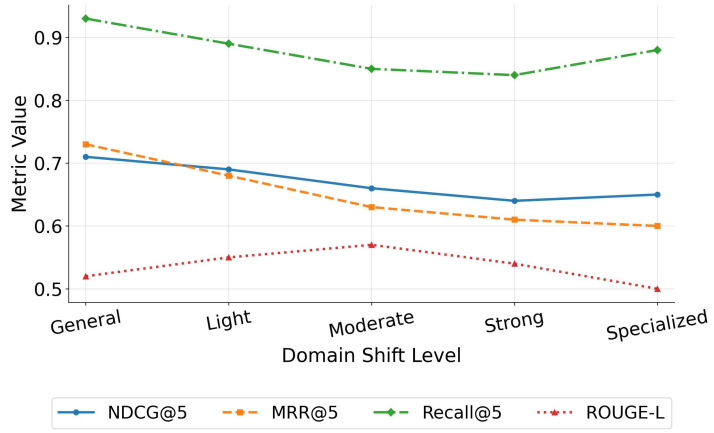


Figure 5.Data Sensitivity Evaluation of Sparse-Dense Dual Retrieval Adaptability under Domain Transfer (General→Specialized)

In the NDCG@5 results, the ranking quality shows a general decline as the domain shifts from general to specialized, but a slight recovery appears in fully specialized settings. This indicates that when the corpus gradually deviates from the original distribution, the model's ability in semantic matching is weakened. Once it fully enters a specialized domain, the adaptability of the dense retrieval path compensates for part of the loss in ranking relevance, reflecting the regulatory role of dual retrieval under semantic shifts.

The change in MRR@5 is more drastic, with a rapid drop during mild shifts followed by stabilization in later stages. This suggests that the ability to capture the first relevant candidate is highly sensitive to distributional shifts, especially when domain features differ greatly from general features. Compared with the NDCG curve, the faster decline in MRR reveals that single candidate localization accuracy is more easily affected in cross-domain scenarios. This also highlights the need for adaptive re-ranking to improve robustness under different distributions.

The Recall@5 results show a decline followed by recovery. Coverage is significantly constrained during moderate shifts but gradually restored in specialized contexts through the compensatory ability of the dense path. This demonstrates that sparse retrieval suffers from insufficient coverage under domain specialization, while dense retrieval can offset this weakness and preserve the completeness of the candidate set. The outcome further confirms the complementary value of sparse–dense fusion in transfer scenarios.

The ROUGE-L trend differs from ranking metrics. It increases during mild to moderate shifts but declines under strong shifts. This indicates that moderate domain specialization helps improve consistency in generation, as diverse evidence enriches text outputs. However, excessive domain differences introduce noise and instability, leading to a decline in generation quality. These results

emphasize the importance of balancing evidence diversity and relevance in cross-domain RAG. They also provide a key motivation for the design of adaptive re-ranking.

Finally, this study tested the data display sensitivity of label sparsity and feedback noise ratio to alignment robustness, and the experimental results are shown in Figure 6.

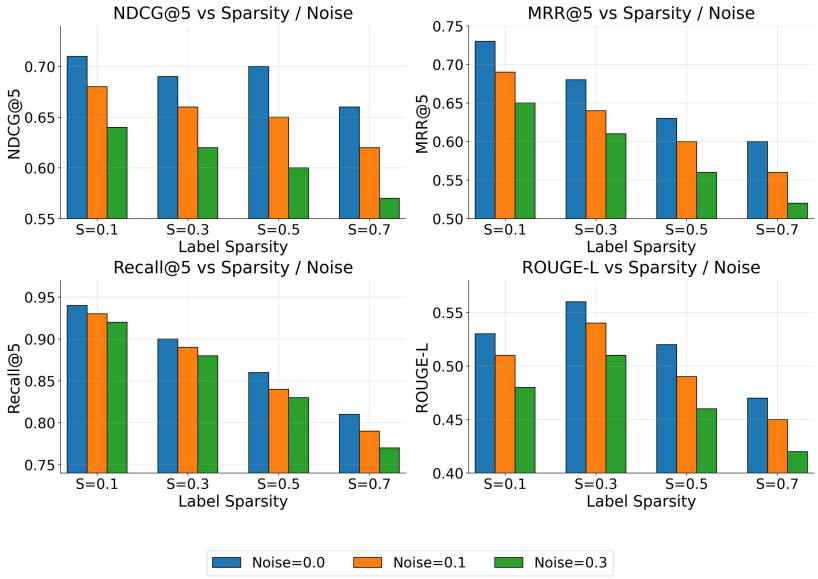


Figure 6.Data display sensitivity test of label sparsity and feedback noise ratio on alignment robustness

In the NDCG@5 results, ranking quality declines as the proportion of sparse labels increases, and the introduction of feedback noise further amplifies this decline. Under low-noise conditions, the model still maintains a degree of robustness. However, in high-noise and high-sparsity settings, ranking performance is significantly weakened. This indicates that the stability of contextual relevance modeling is challenged when evidence quality decreases.

The MRR@5 curve shows an even sharper downward trend, especially in medium-to-high sparsity ranges, where the ability to identify the first relevant candidate is severely affected. This suggests that in environments with insufficient annotations and unstable feedback, the precision of locating the optimal answer declines markedly. Compared with NDCG@5, the MRR@5 results more clearly reveal the importance of label completeness and noise control for alignment robustness under fine-grained accuracy requirements.

The changes in Recall@5 display a decline mainly driven by label sparsity, with feedback noise having a relatively smaller effect. This suggests that coverage largely depends on the proportion of annotations available during training. When sparsity is too high, the probability of missing relevant evidence in the candidate pool increases significantly. Although noise has some effect on coverage, its role is secondary compared with the direct information loss caused by sparsity.

The ROUGE-L results present a non-monotonic pattern. At moderate sparsity levels, generation quality may show a temporary improvement, but it drops rapidly under high sparsity and high noise. This indicates that moderate sparsity and mild noise can increase sample diversity and improve generation consistency. However, when noise and sparsity accumulate excessively, the semantic stability and completeness of the generated text are severely damaged. These findings verify the necessity of adaptive re-ranking in maintaining generation quality under high-uncertainty conditions.

5. Conclusion

This study investigates a sparse – dense dual retrieval fusion and adaptive re-ranking framework, and proposes a new method that balances ranking quality and generation quality. Experimental results across different scenarios show that the method effectively mitigates the conflict between coverage and relevance in retrieval evidence, providing higher-quality contextual support for generative models. By systematically examining hyperparameters, environments, and data sensitivity, the study not only verifies the robustness of the framework but also reveals the adaptation mechanisms of RAG under complex conditions, offering a practical reference paradigm for future research.

At the methodological level, the complementarity of sparse and dense retrieval is fully demonstrated, while the re-ranking mechanism provides an adaptive optimization path for evidence from different sources. This design breaks the limitation of single retrieval paths, enabling RAG systems to maintain stability and generalization in multi-source and dynamic data environments. More importantly, the framework shows how structured modeling and dynamic adjustment can achieve cross-scenario transferability, which lays the foundation for improving the deployment value of large-scale RAG systems.

At the application level, this study offers insights for multiple directions, such as knowledge question answering, intelligent retrieval, dialogue generation, and professional information services. The integration of sparse and dense retrieval not only improves robustness at the retrieval stage but also enhances contextual consistency at the generation stage. This allows models to better handle data noise, limited annotations, and domain shifts in real-world settings. Such capability is critical for building reliable information and interaction systems and may promote the application of RAG technologies in industries such as education, healthcare, law, and finance.

Future research can further explore how to achieve efficient joint optimization of dual retrieval and re-ranking in large-scale distributed environments. It can also integrate privacy protection and federated learning mechanisms to maintain stability under data isolation. In addition, incorporating multimodal evidence could improve generalization and robustness in cross-modal tasks. As RAG continues to expand into diverse application scenarios, the proposed framework has the potential to become an important component of trustworthy intelligent systems and to play a greater role in connecting artificial intelligence technologies with real-world industry practices.

References

[1] 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] 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.

[3] Izacard G, Grave E. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering[C]//International Conference on Learning Representations. 2021.

[4] Yan C, Ma M, Zhang Y, et al. Dual-Path Multimodal Optimal Transport for Composed Image Retrieval[C]//Proceedings of the Asian conference on computer vision. 2024: 1741-1755.

[5] Huang J, Wang M, Cui Y, et al. Layered Query Retrieval: An Adaptive Framework for Retrieval-Augmented Generation in Complex Question Answering for Large Language Models[J]. Applied Sciences, 2024, 14(23): 11014.

[6] 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.

[7] Lu J, Hall K, Ma J, et al. Hyrr: Hybrid infused reranking for passage retrieval[J]. arXiv preprint arXiv:2212.10528, 2022.

[8] Wang Z, Mao Y, Wu W, et al. RAGAS: Automated Evaluation of Retrieval Augmented Generation[C]//Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations. 2024: 150-158.

[9] 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.

[10] Su J, Zhou J P, Zhang Z, et al. Towards more robust retrieval-augmented generation: Evaluating rag under adversarial poisoning attacks[J]. arXiv preprint arXiv:2412.16708, 2024.

[11] Niu T, Joty S, Liu Y, et al. Judgerank: Leveraging large language models for reasoning-intensive reranking[J]. arXiv preprint arXiv:2411.00142, 2024.

[12] 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.

[13] de Jong M, Zemlyanskiy Y, Ainslie J, et al. Fido: Fusion-in-decoder optimized for stronger performance and faster inference[J]. arXiv preprint arXiv:2212.08153, 2022.

[14] Yoon S, Choi E, Kim J, et al. Listt5: Listwise reranking with fusion-in-decoder improves zero-shot retrieval[J]. arXiv preprint arXiv:2402.15838, 2024.