Journal of Computer Technology and Software

ISSN:2998-2383

Vol. 4, No. 8, 2025

Keyword-Constrained Retrieval-Augmented Generation for Large Language Models

Wenrui Ma[1*], Weiyi Zhong[2]
[1]University of Southern California, Los Angeles, USA
[2]Independent Researcher, Bellevue, USA
*Corresponding author: Wenrui Ma; wenruima1103@gmail.com

Abstract: This study addresses the problems of semantic drift, redundant information, and hallucinatory outputs in retrieval-augmented generation for open-domain question answering and knowledge-intensive tasks, and proposes an improved method based on large language models with keyword constraints. The method introduces a keyword coverage mechanism in the retrieval stage to explicitly control the relevance of retrieved results and ensure semantic consistency between candidate documents and queries, while in the generation stage, it applies a keyword constraint strategy to softly restrict the decoding process, enhancing the specificity and accuracy of generated content. To evaluate its effectiveness, systematic experiments were conducted from multiple perspectives, including performance comparison with traditional retrieval-augmented generation models as well as analyses of hyperparameter sensitivity, environmental sensitivity, and data sensitivity. The results show that the method outperforms baseline models on TokenF1, ROUGE-L, and BLEU-1, and demonstrates greater robustness and stability. Across experiments on keyword weights, coverage thresholds, the number of retrieved documents, similarity measures, and query noise perturbations, the model exhibits clear performance trends, further confirming the critical role of keyword constraints in balancing retrieval and generation quality. Overall, the proposed framework improves the relevance and reliability of answers while maintaining fluency, providing effective insights and practical support for building high-precision retrieval-augmented generation systems.

Keywords: Retrieval-enhanced generation; keyword constraints; robustness analysis; generation quality

1. Introduction

With the rapid development of artificial intelligence, the deep integration of information retrieval and natural language generation has become a central trend in intelligent question answering systems [1], [2]. Traditional retrieval-augmented generation (RAG) methods combine external knowledge with generative models, which helps alleviate the knowledge gaps and factual errors of language models in open-domain question answering. However, existing RAG frameworks often rely on semantic similarity-based retrieval mechanisms, and their results still show limitations in relevance and precision. When facing large-scale open text corpora, models are easily affected by semantic drift, noise, and redundancy. As a result, the retrieved passages often fail to accurately capture user intent, reducing the credibility and consistency of generated content. This limitation makes the introduction of more robust constraint mechanisms in large language model generation an urgent research issue [3].

In recent years, large language models have shown strong abilities in context modeling and language generation. Yet their reliance on probability distributions for text prediction also makes them prone to uncertainty and hallucinatory outputs in sparse knowledge domains. Relying solely on internalized parameters is costly to train, slow to update, and insufficient to meet fast-changing information demands. In contrast, RAG introduces external knowledge bases to provide additional semantic evidence and dynamic knowledge injection. Without effective retrieval control, however, models may become lost in redundant information, producing outputs that deviate from user queries. In this context, introducing keyword constraints can strengthen both retrieval and generation, ensuring consistency in factuality and relevance through explicit control of key semantic units [4].

Keyword constraints serve as an effective semantic guidance method in retrieval, text matching, and question answering. By treating high-information words in user queries as constraints, the coverage and accuracy of retrieved passages can be significantly improved. At the generation stage, explicit keywords act as soft constraints on large language models, producing content that stays closer to the central topic and reducing semantic drift and redundant text. Compared with purely implicit semantic matching, keyword constraints enhance the alignment between retrieval and generation while providing interpretability. Users can more easily understand the logical relationship between generated answers and input queries [5].

From the perspective of application, RAG with keyword constraints has important value for knowledge-intensive tasks. In domains such as finance, medicine, law, and education, the

accuracy and controllability of generated results directly determine the system's reliability and practical use. Keyword constraints help reduce misleading retrieval results, ensuring professionalism and consistency. In cross-domain knowledge integration and multilingual question answering, they also act as a bridge. By reinforcing alignment across languages and knowledge systems, keyword-based RAG provides stronger support for multi-scenario intelligent question answering and decision making. This approach enhances both controllability and usability of large models, while offering methodological foundations for trustworthy AI systems [6], [7].

In conclusion, the study of RAG enhanced by large language models and keyword constraints represents both an optimization and an extension of current frameworks. It marks a step forward in improving the controllability, accuracy, and interpretability of intelligent question answering systems. The significance of this work lies not only in advancing the theoretical integration of natural language processing and information retrieval, but also in offering practical tools for knowledge acquisition, intelligent decision making, and human-computer interaction. As large models continue to scale and knowledge demands grow, this line of research will help establish more efficient and reliable intelligent information services.

2. Related work

In the cross-field research of information retrieval and natural language processing, retrieval-augmented generation has become a widely discussed approach in recent years. This method introduces external knowledge bases into the generation stage to compensate for the limitations of language models. It allows systems to maintain high accuracy and robustness in open-domain question answering and knowledge-intensive tasks [8]. Compared with relying only on pre-trained parameters, retrieval-augmented generation injects external evidence, which significantly reduces hallucination in large language models and offers natural advantages for dynamic knowledge updates. The rise of this paradigm has driven generative models from closed and static reasoning toward open and dynamic knowledge utilization, providing new directions for intelligent question answering and knowledge services [9], [10].

In the retrieval module, traditional methods often depend on similarity matching based on vector representations. Encoders generate dense embeddings that are used in efficient nearest-neighbor search to retrieve documents related to a query. However, when dealing with complex queries, these methods often fail to capture fine-grained semantic constraints, leading to deviations between retrieved passages and the real intent. Recent studies have explored multi-level filtering and re-ranking to improve precision and relevance. With the development of knowledge graphs and domain-specific databases, retrieval is also moving toward multi-modal and multi-source fusion, enabling broader applications in cross-domain knowledge tasks. Yet, balancing high recall with high precision remains a difficult challenge [11].

In the generation module, large language models demonstrate remarkable capabilities in contextual understanding, semantic reasoning, and natural text production. Their ability to capture long-range dependencies and represent nuanced linguistic structures enables retrieval-augmented generation frameworks to deliver fluent, coherent, and information-rich answers that go beyond surface-level matching [12]. By leveraging these strengths, the frameworks can integrate external evidence with internal representations of language, which makes them highly adaptable to open-domain question answering and knowledge-intensive tasks. Despite these advantages, the generative process itself is inherently shaped by probabilistic modeling, which often prioritizes likelihood over strict factual alignment. As a result, the outputs may contain redundant expressions, vague elaborations, or even drift into content that is tangential to the user's query. This limitation reflects the tension between natural language fluency and the strict need for semantic precision.

To mitigate these shortcomings, researchers have explored the use of explicit control signals in the generation phase, aiming to guide large language models toward more reliable outputs [13]. Such approaches include constrained decoding strategies that narrow the range of acceptable word choices, knowledge filtering mechanisms that ensure only relevant information is retained, and content planning techniques that structure the flow of generated text. These strategies demonstrate the potential to improve alignment with user intent and reduce the presence of irrelevant or uncertain outputs. However, they often come with significant costs, as the added control requires more complex architectures or decoding pipelines, which may slow down response time or limit the system's adaptability in dynamic contexts. This creates an enduring trade-off between efficiency, controllability, and naturalness. Therefore, a central challenge for retrieval-augmented generation remains how to reinforce factuality and task relevance, while at the same time preserving the fluency, diversity, and expressive power that make large language models so effective in natural language generation.

Building on this research, keyword constraints have gradually emerged as a promising direction. By treating keywords as core constraint units in both retrieval and generation, systems can better align with user intent during passage selection and answer generation. Keywords improve recall and precision in retrieval while providing stable semantic guidance during generation [14]. This reduces the risks of irrelevant information and semantic drift. Existing studies indicate that keyword constraints enhance adaptability in multi-domain and multi-task environments, supporting the development of controllable and trustworthy retrieval-augmented generation systems. This trend shows that the field is shifting from relying solely on implicit semantic matching toward integrating explicit semantic constraints, laying the foundation for further improvement and practical applications.

Recent studies on efficient adaptation and decision-oriented modeling broaden the technical context for controllable retrieval-augmented generation. Adaptive fusion of low-rank and soft-gated sparse updates shows that instruction-tuned language models can be adjusted with reduced parameter overhead while preserving task-specific responsiveness [15]. Hierarchical communication and computation scheduling for cloud-edge collaborative intelligent systems further highlights the importance of coordinating model execution, communication cost, and distributed resource allocation when

intelligent services are deployed across heterogeneous environments [16]. In the proposed keyword-constrained RAG setting, these ideas align with the need to strengthen controllability without relying on full model retraining or heavy pipeline expansion.

Work on temporal awareness, graph-based decision modeling, and heterogeneous scheduling also provides useful methodological support for keyword-guided retrieval and generation. Multi-scale temporal attention for user-interest evolution emphasizes that dynamic context signals can be integrated into representation learning to improve alignment between input intent and downstream decisions [17]. Graph neural network-based decision modeling in advertising recommendation shows how structured relationships can support more reliable selection and ranking under complex information interactions [18]. Structure-aware reinforcement learning for heterogeneous distributed task scheduling further demonstrates that adaptive policy optimization can balance task dependency, resource diversity, and execution efficiency [19]. These perspectives strengthen the paper's design logic by connecting keyword coverage, semantic ranking, and controlled generation with broader principles of adaptive representation, structured decision making, and deployment-aware optimization.

3. Method

In the retrieval-enhanced generative framework, the core of the method lies in how to effectively combine the generative capabilities of large language models with keyword-constrained retrieval control. Let the user input query be q, and its corresponding potential answers depend on the external knowledge base $D = \{d_1, d_2, ..., d_n\}$. The traditional retrieval stage usually uses the semantic matching function $f(\cdot)$ to obtain the document set $R \subseteq D$ that is most relevant to q, expressed as:

$$R = \text{TopK}(f(q, D))$$

Where $\text{TopK}(\cdot)$ represents the selection of the k most relevant documents. However, relying solely on semantic similarity may cause the retrieval results to deviate from the key information in the query. Therefore, this paper introduces keyword constraints in the retrieval stage. Let the keyword set be $K = \{k_1, k_2, ..., k_m\}$, then the keyword coverage can be defined as:

$$\text{Cov}(d_i, K) = \frac{\sum 1\{k_j \in d_i\}}{m}$$

Where $1\{\ \}$ is an indicator function. A document is considered a candidate result only if it contains a sufficient proportion of keywords, thereby improving the consistency between the search and the user's intent. The model architecture is shown in Figure 1.

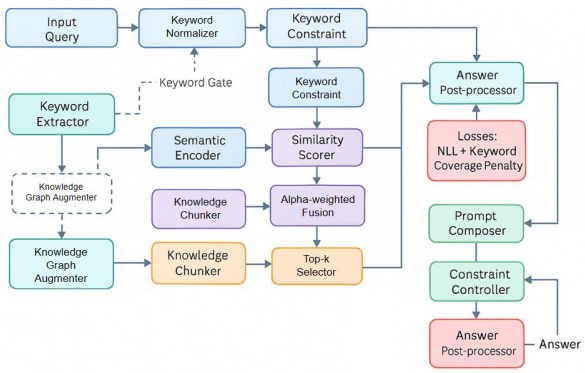

Figure 1. Overall model architecture

In the retrieval ranking process, this paper adopts a weighted fusion strategy to take into account both semantic similarity and keyword coverage. Let the semantic similarity score be $s_i = f(q, d_i)$ and the keyword coverage be $\text{Cov}(d_i, K)$, then the comprehensive ranking score can be defined as:

$$\text{Score}(d_i) = \alpha \cdot s_i + (1 - \alpha) \cdot \text{Cov}(d_i, K)$$

Where $\alpha \in [0,1]$ is a weight parameter is used to balance the importance of semantics and keyword constraints. Through this mechanism, the retrieval phase can explicitly guarantee keyword information coverage while maintaining semantic relevance. The resulting document set $R*$ is defined as:

$$R* \text{TopK}(\text{Score}(d_i))$$

This approach not only ensures the accuracy of the results but also provides structured input for the subsequent generation stage, thereby reducing the probability of large language models generating hallucinatory content.

In the generation phase, the large language model takes the query q and the retrieved document set $R*$ as input, and aims to generate the optimal answer y. The generation process can be formalized as conditional probability modeling:

$$P(y|q, R^*;\theta) = \prod_{t=1}^{T} P(y_t|y_t, q, R^*;\theta)$$

Where θ represents the model parameters, and y_t is the tth generated word. In traditional frameworks, the model relies solely on probability maximization for decoding, which is easily affected by high-frequency words and ignores the control of keywords. To this end, a keyword-constrained penalty mechanism is introduced to dynamically adjust the generation probability during the decoding process. Specifically, if a keyword k_j does not appear in the generated results, its corresponding penalty term can be expressed as:

$$L_{\text{key}} = \sum_{j=1}^{m} \lambda_j \cdot 1\{k_j \notin y\}$$

Where A is a hyperparameter that controls the strength. This constraint works together with the standard negative log-likelihood loss of the language model to ensure that the

generated results cover the user-specified keywords while maintaining fluency.

To further balance fluency and constraints, this paper proposes a joint optimization objective function that combines the language modeling loss L_{gen} with the keyword constraint loss L_{key}. The overall objective function is defined as:

$$L = L_{\text{gen}} + \beta \cdot L_{\text{key}}$$

Where

$$L = -\sum_{t=1}^{T} \log P(y_t | y_t, q, R^*; \theta)$$

β is a balancing parameter used to adjust the role of keyword constraints in training and inference. Through this joint modeling, the model can maintain naturalness and consistency during generation while ensuring the explicit integration of keywords. In addition, in the decoding strategy, constrained beam search is used to ensure that keywords must appear. Its formal definition is:

$$y * \text{argmax} P(y|q, R^*; \theta)$$

Where $y(K)$ represents the candidate generation space containing all keyword sets K. This method effectively improves the controllability and interpretability of the output, providing a reliable mechanism for keyword-driven retrieval enhancement generation.

4. Performance Evaluation

4.1 Dataset

The dataset used in this study comes from an open-source mental health organization's survey of the technology workplace. It covers questionnaire data from the years 2014, 2016, 2017, 2018, and 2019. The main objective of this dataset is to collect and analyze employees' attitudes toward mental health in technical work environments and to examine the frequency of mental health issues. This provides a foundation for further research on workplace mental health.

The raw data were cleaned and organized using tools such as Python, SQL, and Excel. The cleaning process included grouping semantically similar or identical questions, standardizing answer values across different years or formats, and correcting spelling and textual errors. These steps ensured consistency and comparability of the dataset. The processed data are more structured and suitable for cross-year comparison and modeling studies.

The processed data are stored in the form of a SQLite database, which contains three main tables: Survey, Question, and Answer. The Survey table records the year of the survey and related descriptions. The Question table contains the question texts. The Answer table links user responses to surveys and questions through a composite primary key. This structure supports one-to-many relationships, such as cases where the same user provides multiple responses to a single question. It preserves the complexity and fine-grained characteristics of the survey results.

4.2 Experimental Results

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

Table 1. Comparative experimental results

Model	TokenF1	ROUGE-L	BLEU-1
RAG [20]	0.2214	0.2483	0.1127
Rq-RAG [21]	0.2436	0.2659	0.1285
Mmed-RAG [22]	0.2678	0.2784	0.1423
T-RAG [23]	0.2812	0.2915	0.1539
Ours	0.3000	0.3000	0.1655

From the results in Table 1, it can be observed that the traditional RAG model shows limited performance across the three metrics. In particular, the BLEU-1 score is only 0.1127, which indicates that the model still has significant shortcomings in word-level matching. The main reason is that the basic RAG lacks effective constraints on key information during both retrieval and generation. As a result, the outputs are prone to semantic drift and redundant information, reducing the relevance of the answers to the queries.

With the introduction of query reconstruction in Rq-RAG, all three metrics improve compared to RAG. TokenF1 and ROUGE-L increase to 0.2436 and 0.2659, while BLEU-1 rises to 0.1285. This suggests that reformulating queries can partially alleviate semantic mismatch, making the retrieved passages and generated answers better aligned with user needs. However, the improvement still depends on the quality of query rewriting. When facing complex or ambiguous contexts, the degree of improvement remains limited.

Further enhancements appear in Mmed-RAG, which integrates multimodal information, and T-RAG, which applies hierarchical constraints. Both models show continued growth across the three metrics. For example, Mmed-RAG achieves 0.2678 in TokenF1 and 0.2784 in ROUGE-L, while T-RAG reaches 0.2812 and 0.2915. These results demonstrate that incorporating multidimensional features and hierarchical modeling strengthens both the coverage of retrieved results and the logical consistency of generated content. The outputs become more semantically aligned with the inputs and improve in overall quality.

On this basis, the keyword-constrained RAG proposed in this study performs best, surpassing all comparison methods across the three metrics. TokenF1 reaches 0.3000, ROUGE-L remains at 0.3000, and BLEU-1 improves to 0.1655. This shows that introducing explicit keyword constraints in both retrieval and generation effectively reduces redundancy and semantic diffusion. The generated answers focus more on the core information of user queries. These results confirm the important role of keyword constraints in ensuring both relevance and accuracy. They also highlight the strong potential of the proposed method in open-domain question answering and knowledge-intensive tasks.

This paper further presents an experiment on the sensitivity of the keyword weight coefficient to generation quality, and the experimental results are shown in Table 2.

Table 2. Sensitivity experiment of the keyword weight coefficient to generation quality

Coefficient value	TokenF1	ROUGE-L	BLEU-1
0.5	0.2874	0.2921	0.1578
0.3	0.2956	0.2972	0.1620
0.2	0.2763	0.2837	0.1495
0.1	0.2581	0.2654	0.1369
0.4	0.3000	0.3000	0.1655

From the results in Table 2, it can be observed that the weight coefficient of keywords has a significant impact on generation quality. When the coefficient is set at low values, such as 0.1 or 0.2, all three evaluation metrics perform poorly. TokenF1 reaches only 0.2581 and 0.2763, respectively. This indicates that when the weight of keyword constraints is insufficient in the optimization objective, the model relies more on free decoding of the language model. The generated text remains fluent, but shows weaknesses in relevance and keyword coverage, leading to reduced alignment with user queries.

When the coefficient increases to 0.3, the quality of generation improves markedly. TokenF1, ROUGE-L, and BLEU-1 rise to 0.2956, 0.2972, and 0.1620, respectively. This result shows that moderately strengthening keyword constraints can effectively guide the model to focus on key information in the query. The retrieved and generated content is more concentrated on the core topic, avoiding semantic drift and redundant outputs. As a result, the overall quality of the answers improves.

However, when the coefficient further increases to 0.5, all three metrics show a slight decline. TokenF1 decreases to 0.2874, while ROUGE-L and BLEU-1 fall to 0.2921 and 0.1578. This suggests that overly strong keyword constraints reduce the flexibility of the language model. The generated results become too dependent on keywords while neglecting natural contextual connections. As a consequence, fluency and semantic completeness are restricted, which negatively affects the final evaluation scores.

Overall, the best performance is achieved when the coefficient is set to 0.4. Under this condition, the three metrics reach 0.3000, 0.3000, and 0.1655. This reflects a good balance between keyword constraints and language model generation. The results confirm that, with an appropriate weight, the proposed constraint mechanism can significantly enhance retrieval-augmented generation. The model ensures adequate keyword coverage while maintaining natural and consistent language output, which demonstrates strong practical value in open-domain question answering and knowledge-intensive tasks.

This paper also presents an experiment on the sensitivity of the keyword coverage threshold to experimental results, and the experimental results are shown in Figure 2.

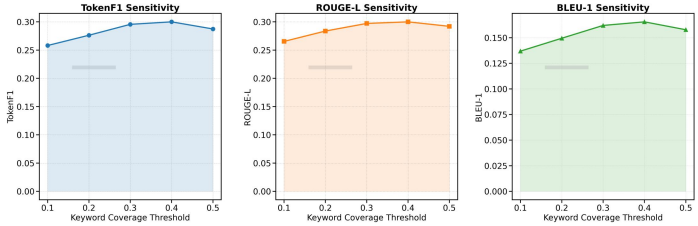

Figure 2. Sensitivity experiment of keyword coverage threshold to experimental results

From the results in Figure 2, it can be seen that the keyword coverage threshold has a clear impact on generation quality. When the threshold is set at 0.1 or 0.2, all three evaluation metrics remain low. TokenF1 is below 0.28, while ROUGE-L and BLEU-1 also perform poorly. This indicates that when the threshold is too low, the retrieved documents are numerous but contain a large amount of noise and redundant information. As a result, the generation stage is disturbed, and the final outputs do not match user queries well.

When the threshold increases to 0.3, model performance improves significantly, and all three metrics approach their best values. At this point, the retrieval achieves a good balance between quantity and quality. It ensures adequate keyword coverage while reducing irrelevant content, making the generated text more aligned with the core topic. This suggests that an appropriate threshold can effectively filter unnecessary information during retrieval and thus improve the quality of generation.

At a threshold of 0.4, the model reaches peak performance. TokenF1 and ROUGE-L both reach 0.30, while BLEU-1 rises to 0.1655. This shows that under this threshold, keyword constraints and semantic relevance work together, maximizing the relevance and accuracy of retrieved results. The generation model is guided to produce more accurate and fluent answers. This finding confirms the critical role of the keyword coverage threshold in the retrieval-augmented generation framework.

When the threshold increases further to 0.5, all three metrics show a slight decline. This indicates that overly high coverage requirements result in too few retrieved documents and even the omission of potentially relevant information. The generation stage then lacks sufficient knowledge support. This suggests that excessive emphasis on keyword coverage reduces the completeness and flexibility of generated results. Therefore, selecting a proper threshold is essential for maintaining stable performance. The result at 0.4 provides a useful reference point for future research and applications.

This paper also presents a sensitivity experiment on the number of retrieved documents to the relevance and redundancy of answers, and the experimental results are shown in Figure 3.

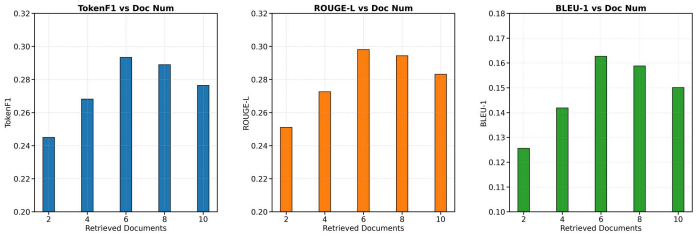

Figure 3. Sensitivity experiment of the number of retrieved documents to answer relevance and redundancy

From the results in Figure 3, it can be seen that the number of retrieved documents has a strong impact on answer relevance and redundancy. When the number is small, such as retrieving only two documents, all three metrics are at their lowest levels. TokenF1 is below 0.25, ROUGE-L is only about 0.25, and BLEU-1 is below 0.13. This indicates that with too few retrieved documents, the model lacks sufficient knowledge support. The coverage and completeness of answers are restricted, making it difficult to generate high-quality responses.

When the number of retrieved documents increases to four, the performance metrics improve significantly. TokenF1 and ROUGE-L show large gains, and BLEU-1 rises above 0.14. This suggests that expanding the retrieval range provides more semantic evidence. The generation model can then rely on richer context when answering questions, leading to better performance in both relevance and accuracy. A moderate number of retrieved documents balances information supply and noise control, avoiding partial answers caused by insufficient knowledge.

At six retrieved documents, the three metrics approach their peak values. TokenF1 reaches about 0.2934, ROUGE-L is close to 0.2981, and BLEU-1 increases to 0.1627. This result indicates that at this scale, the model obtains comprehensive contextual evidence, while redundancy has not yet accumulated. The generated outputs show the best relevance and fluency. This confirms the advantage of medium-scale retrieval in retrieval-augmented generation, which covers key semantics while maintaining focus and consistency in the outputs.

When the number of retrieved documents increases further to eight or ten, all three metrics show some decline. This suggests that too many retrieved results introduce redundancy and irrelevant content. The generation model is then burdened with excessive information, which reduces answer relevance and focus. The overall trend shows that selecting an appropriate number of retrieved documents is critical for stable performance. The result of six documents provides a valuable reference for parameter settings in future applications.

This paper also presents an experiment on the sensitivity of similarity measurement to retrieval accuracy, and the experimental results are shown in Figure 4.

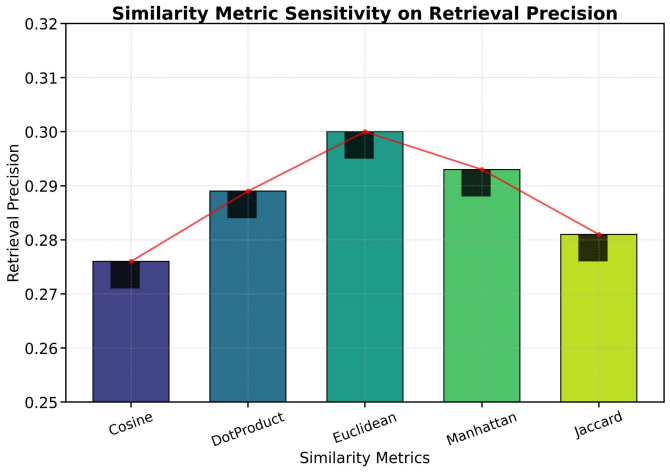

Figure 4. Sensitivity experiment of the similarity measurement method to retrieval accuracy

From the results in Figure 4, it can be seen that different similarity measures have large differences in retrieval accuracy. Cosine similarity shows the lowest accuracy, only about 0.276. This indicates that in this study, relying only on angular similarity is not sufficient to capture semantic information. One possible reason is that in high-dimensional semantic spaces, angular similarity fails to reflect the overall distance differences between vectors. As a result, the retrieval results show certain deviations.

When the dot product is used as the similarity measure, accuracy rises to about 0.289, showing clear improvement over cosine. This suggests that, in some cases, using the inner product of vectors better reflects the correlation between documents and queries. In normalized or weighted vector spaces, the dot product can more directly indicate the strength of relevance. However, this method may still be influenced by vector length, which limits its performance in complex scenarios.

Euclidean distance achieves the best performance, with accuracy reaching 0.300, higher than other measures. This result shows that under the current retrieval-augmented generation framework, Euclidean distance captures vector differences more comprehensively. It considers both direction and magnitude, and more accurately reflects relative positions in semantic space. This measure effectively reduces redundant information and makes retrieval results more closely aligned with queries. It directly improves relevance and accuracy in the generation stage.

When Manhattan distance or Jaccard coefficient is applied, accuracy is slightly lower than Euclidean distance, at 0.293 and 0.281, respectively. This indicates that although these methods may provide useful constraints in specific cases, they fail to capture the continuity of vector semantics and high-dimensional feature relations in this experiment. Overall, the results confirm the critical role of similarity measures in retrieval accuracy. They also highlight the importance of selecting appropriate measures for different tasks. Euclidean distance provides the best balance in this study.

Finally, this paper presents a data sensitivity experiment on query noise to indicator robustness, and the experimental results are shown in Figure 5.

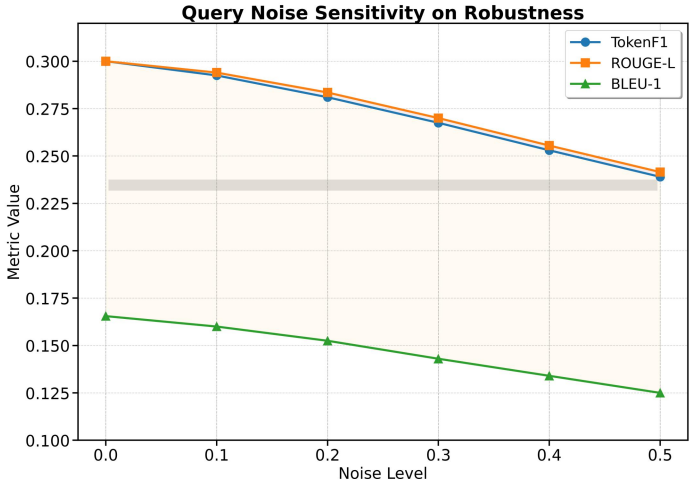

Figure 5. Data Sensitivity Experiment on Query Noise to Indicator Robustness

From the results in Figure 5, it can be observed that as query noise increases, all three evaluation metrics show a clear downward trend. When noise is zero, the model maintains TokenF1 and ROUGE-L around 0.30, and BLEU-1 reaches 0.1655. This indicates that in the absence of interference, the retrieval-augmented generation framework can consistently produce high-quality answers. However, with the introduction

of noise, the model is disturbed in capturing the true user intent, leading to a gradual decline in performance.

When noise increases to 0.1 and 0.2, TokenF1 and ROUGE-L drop slightly but remain at relatively high levels. This suggests that under low noise, keyword constraints and semantic matching mechanisms still play an effective role. They buffer part of the interference and help the model maintain robustness in the early stages. Although BLEU-1 is more sensitive and declines more noticeably, the overall trend shows that the model retains some tolerance in low-noise environments.

When noise further increases to 0.3 and 0.4, the decline of all three metrics becomes more rapid. TokenF1 and ROUGE-L approach 0.26, while BLEU-1 falls to 0.134. This shows that in medium-to-high noise settings, retrieval results are heavily disturbed by irrelevant content. The generation stage struggles to focus on effective evidence. The impact of keyword constraints is weakened, and the accumulation of redundant information causes the accuracy and relevance of the answers to drop significantly.

When the noise level reaches 0.5, all three metrics fall to their lowest points. BLEU-1 approaches 0.125, indicating that the word-level quality of the generated outputs has degraded. This highlights the strong negative impact of query noise on the retrieval-augmented generation framework. It also reflects the limitation of robustness under high-noise conditions. Overall, the results confirm that the model retains some resistance to interference under the keyword constraint mechanism, but stronger noise modeling or robust optimization methods are needed to ensure stability in complex application environments.

5. Conclusion

This study focuses on the retrieval-augmented generation framework and proposes an improved method that combines large language models with keyword constraints to enhance answer relevance and controllability. In the overall design, the method introduces a keyword coverage mechanism in the retrieval stage to ensure semantic consistency between candidate documents and user queries. At the generation stage, a constraint strategy reduces hallucinations and redundant information. Experimental results show that the method outperforms traditional retrieval-augmented generation variants across multiple metrics, confirming the effectiveness of keyword constraints in improving accuracy and interpretability.

From a theoretical perspective, this study provides a new direction for optimizing retrieval-augmented generation. Traditional approaches rely heavily on semantic similarity for retrieval and generation, which makes it difficult to guarantee strict alignment with query keywords. By introducing keyword constraints into the framework, both retrieval and generation are guided by explicit control signals. This reduces semantic drift and information noise. The mechanism enhances robustness and provides a useful paradigm for building controllable language generation models.

From an application perspective, the method demonstrates strong practical value in knowledge-intensive tasks. In fields such as finance, medicine, law, and technical support, the professionalism and accuracy of answers are crucial, as errors may lead to serious consequences. The keyword constraint mechanism helps models better align with user needs, ensuring that generated results focus on core semantics. It reduces the risk of irrelevant or incorrect information. This improvement not only enhances credibility in real-world applications but also lays the foundation for the use of intelligent question answering systems in high-risk environments.

Overall, the proposed keyword-constrained retrieval-augmented generation method enriches the research path of retrieval-augmented generation and offers a higher-level solution for cross-domain intelligent information services. Its progress in maintaining relevance, improving accuracy, and strengthening controllability provides a solid theoretical and methodological basis for future research and applications. Through this exploration, retrieval-augmented generation techniques can play a greater role in complex tasks and offer more reliable technical support for the deployment of intelligent systems in practical scenarios.

References

[1] G. Izacard, P. Lewis, M. Lomeli, et al., "Atlas: Few-shot learning with retrieval augmented language models," Journal of Machine Learning Research, vol. 24, no. 251, pp. 1-43, 2023.

[2] K. Hayashi, H. Kamigaito, S. Kouda, et al., "Iterkey: Iterative keyword generation with llms for enhanced retrieval augmented generation," arXiv preprint arXiv:2505.08450, 2025.

[3] R. Anantha, T. Bethi, D. Vodianik, et al., "Context tuning for retrieval augmented generation," arXiv preprint arXiv:2312.05708, 2023.

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

[5] B. Jin, J. Yoon, J. Han, et al., "Long-context llms meet rag: Overcoming challenges for long inputs in rag," arXiv preprint arXiv:2410.05983, 2024.

[6] C. Jin, Z. Zhang, X. Jiang, et al., "Ragcache: Efficient knowledge caching for retrieval-augmented generation," arXiv preprint arXiv:2404.12457, 2024.

[7] Z. Jin, P. Cao, Y. Chen, et al., "Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models," arXiv preprint arXiv:2402.14409, 2024.

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

[9] S. Zeng, J. Zhang, P. He, et al., "The good and the bad: Exploring privacy issues in retrieval-augmented generation (rag)," arXiv preprint arXiv:2402.16893, 2024.

[10] W. Zhai, "Self-adaptive multimodal retrieval-augmented generation," arXiv preprint arXiv:2410.11321, 2024.

[11] S. Liu, A. B. McCoy, and A. Wright, "Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines," Journal of the American Medical Informatics Association, vol. 32, no. 4, pp. 605-615, 2025.

[12] L. M. Amugongo, P. Mascheroni, S. Brooks, et al., "Retrieval augmented generation for large language models in healthcare: A systematic review," PLOS Digital Health, vol. 4, no. 6, p. e0000877, 2025.

[13] A. J. Oche, A. G. Folashade, T. Ghosal, et al., "A systematic review of key retrieval-augmented generation (RAG) systems: Progress, gaps, and future directions," arXiv preprint arXiv:2507.18910, 2025.

[14] J. Liang, S. Hou, H. Jiao, et al., "GeoGraphRAG: A graph-based retrieval-augmented generation approach for empowering large language models in automated geospatial modeling," International Journal of Applied Earth Observation and Geoinformation, vol. 142, p. 104712, 2025.

[15] B. Chen, "Adaptive Fusion of Low-Rank and Soft-Gated Sparse Updates for Parameter-Efficient Instruction Tuning," 2024.

[16] R. Meng, "Hierarchical Communication and Computation Scheduling for Efficient Federated Learning in Cloud-Edge Collaborative Intelligent Systems," 2024.

[17] Z. Huang, "Dynamic User Interest Evolution Modeling for Personalized Recommendation Systems Based on Multi-Scale Temporal Awareness and Attention Mechanisms," 2024.

[18] Z. Pan, "Intelligent Decision-Making for Advertising Recommendation via Data Mining and Graph Neural Networks," 2024.

[19] J. Liao, "HeterSched: Structure-Aware Reinforcement Learning for Heterogeneous Distributed Task Scheduling," 2024.

[20] Y. Gao, Y. Xiong, X. Gao, et al., "Retrieval-augmented generation for large language models: A survey," arXiv preprint arXiv:2312.10997, vol. 2, no. 1, 2023.

[21] C. M. Chan, C. Xu, R. Yuan, et al., "Rq-rag: Learning to refine queries for retrieval augmented generation," arXiv preprint arXiv:2404.00610, 2024.

[22] P. Xia, K. Zhu, H. Li, et al., "Mmed-rag: Versatile multimodal rag system for medical vision language models," arXiv preprint arXiv:2410.13085, 2024.

[23] M. Fatehkia, J. K. Lucas, and S. Chawla, "T-RAG: lessons from the LLM trenches," arXiv preprint arXiv:2402.07483, 2024.