Artificial Intelligence and Computing Innovations

Intelligence Science Publishing

https://www.ispublishing.com

ISSN: 3070-4944

Vo.4, No.1, 2024

Retrieval-Augmented Long-Text Reasoning with Semantic Conflict Detection and Cross-Paragraph Evidence Integration

Shao-yu Huang[1]*

[1]Duke University, Durham, USA

**Corresponding author: syhuang1201@gmail.com*

Abstract: This study addresses common challenges in long-text reasoning, including retrieval noise, dispersed evidence, semantic drift, and inconsistent generation, and proposes a retrieval-augmented generation framework for long-text understanding. The framework performs hierarchical semantic retrieval, cross-paragraph evidence graph modeling, semantic conflict detection and calibration, and evidence-guided generation to extract key segments from large-scale long texts, correct semantic relations, and produce reasoning outputs that align closely with factual content. In the retrieval stage, the model maps queries into a multi-scale semantic space and applies structured candidate filtering to improve cross-document evidence recall. In the evidence construction stage, graph structures represent semantic associations and potential conflicts across paragraphs, forming a unified semantic center that constrains subsequent generation. In the generation stage, evidence-aware attention and conditional control ensure that outputs rely strictly on real paragraph content, reducing hallucinations and enhancing the logical consistency of the reasoning chain. The experiments evaluate the model across answer accuracy, reasoning consistency, factual faithfulness, and cross-paragraph structural integration, and conduct sensitivity analyses on learning rate, context window length, retrieval candidate size, and training data scale. The results show that the framework achieves stable advantages across multiple metrics and generates high-quality reasoning outputs in complex long-text environments. Overall, this study provides a complete technical pathway that covers retrieval, integration, and generation, offering an effective solution for multi-source and multi-paragraph long-text tasks and establishing a methodological foundation for reliable long-text reasoning in knowledge-intensive application scenarios.

Keywords: Long text reasoning; evidence calibration; structured retrieval; condition generation

1. Introduction

In the context of growing demand for complex knowledge-intensive tasks, long-text reasoning has become one of the most challenging core capabilities for large language model applications. Real-world reasoning targets often consist of lengthy materials that span paragraphs, sections, or even multiple documents. These materials contain sparse logical evidence, dispersed semantic fragments, and heterogeneous information from many sources. Traditional end-to-end generation models perform well on general dialogue and short-text tasks. However, they struggle with extremely long texts, have uneven information density, and contain complex reasoning chains [1]. They often suffer from attention degradation, semantic loss, and broken reasoning paths. Although the context length of large models continues to increase, it still cannot cover the full semantic space of legal judgments, scientific papers, technical specifications, or medical reports. This leads to incomplete reasoning, inconsistent citations, or outputs that deviate from the factual content. These problems drive the need for more efficient long-text understanding frameworks [2].

Retrieval-augmented generation has become an important direction for mitigating the long-context problem. It allows models to dynamically access external knowledge bases or document collections and compensates for the limits of parameterized memory. However, when conventional RAG frameworks are applied directly to long-text reasoning, new bottlenecks emerge. Key evidence in long documents often appears in a nonlinear manner across many paragraphs. Simple text slicing or vector similarity retrieval can easily miss important links. It may also introduce irrelevant fragments that have high surface similarity. This can cause semantic drift or confusion in the reasoning chain. At the same time, the hierarchical structure, cross-paragraph logic, and section-level semantic dependencies cannot be captured by traditional retrieval methods. As a result, the generation process lacks a stable semantic foundation when integrating evidence [3].

In this context, retrieval-augmented techniques for long-text reasoning must address not only the amount of information but also its structure. Long-text reasoning is not merely locating textual fragments. It requires building reasoning chains that span paragraphs, topics, and semantic levels. The model must locate key evidence, detect semantic conflicts, and integrate logical dependencies across sections. It must also maintain consistency in the final output. Therefore, retrieval strategies need higher semantic sensitivity. They must understand the internal chapter structure, logical nodes, and topic transitions of long documents. The generation process also requires strict evidence constraints and semantic consistency control. This helps avoid distorted reasoning caused by retrieval noise or conflicting evidence [4, 5].

As the demand for long-text reasoning grows across many fields, robust reasoning frameworks become increasingly important. In legal review, financial risk assessment, medical diagnosis, scientific literature analysis, and compliance verification, an accurate understanding of long documents is essential for trustworthy decisions [6]. In policy or regulatory documents, relevant clauses are often scattered across chapters, and models must connect conditions and constraints. In medical reports, symptoms, test results, and diagnostic bases appear across different sections, and models must form a precise reasoning path. In scientific papers, the model must extract information from the background, methods, data descriptions, and conclusions to build a complete argument chain. Traditional language models and simple retrieval approaches cannot meet the needs of long, cross-structure reasoning. Retrieval-augmented frameworks with deep semantic integration offer a more suitable solution [7].

Overall, retrieval-augmented generation frameworks for long-text reasoning have significant research value. They respond to the core challenge that large models struggle with long contexts. They expand reasoning coverage by dynamically leveraging external knowledge and allow the model to handle information beyond its parameter capacity and context window. They also provide scalable, interpretable, and consistent reasoning abilities for knowledge-intensive applications. This supports the progress from shallow generation to deeper reasoning and from surface imitation to structured understanding. As the

demand for complex textual processing continues to grow in economic, social, and technological systems, building precise, robust, and controllable long-text reasoning frameworks holds both academic value and practical importance. It also contributes to the foundations of future intelligent decision-making systems.

2. Proposed Framework

2.1 Overall Framework

This study constructs a retrieval-enhanced generative framework for long-text reasoning. Through steps such as query semantic modeling, hierarchical retrieval, conflict detection, and evidence integration, a unified reasoning chain is formed from query to final generation. Given input text X and query q, the query is first mapped to a high-dimensional semantic space:

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

Then, the most relevant candidate segments are retrieved from the long document slice set h_q, and the relevance is calculated using a semantic matching function:

$$s_i=sim(h_q,h_{d_i}) \tag{2}$$

The system further constructs a global evidence graph across paragraphs to unify semantic signals from different sources:

$$G=(V,\varepsilon),\ \ V=\{d_i\} \tag{3}$$

Here, edge weight ε represents the semantic dependency across segments. Finally, the integrated evidence information is injected into the conditional generation distribution of the language model:

$$p(y|q,G)=\prod_t p(y_t|y_{<t},q,G) \tag{4}$$

The overall framework aims to establish a structured flow of evidence to ensure the integrity and consistency of the reasoning process in long texts. It organizes the retrieval, alignment, and integration of multi-source information into a coherent pipeline, allowing the model to progressively refine the semantic signals required for accurate reasoning. By transforming scattered textual cues into a unified and well-formed evidence chain, the framework provides a stable foundation for handling complex, cross-paragraph dependencies. Its overall model architecture is shown in Figure 1.

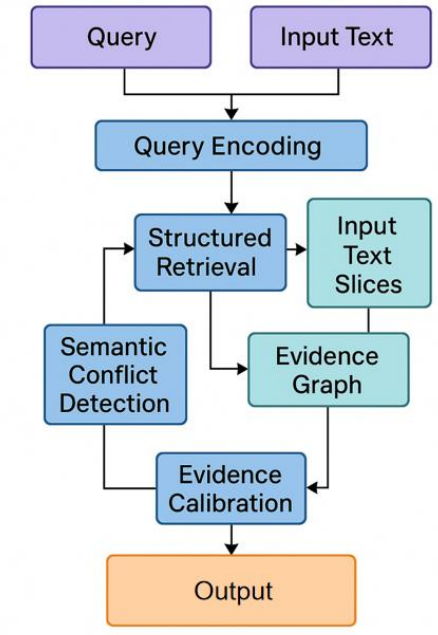

Figure 1. Overall model architecture diagram

2.2 Query Encoding and Structured Retrieval

To accommodate the multi-granular semantic structure within long texts, this study employs a multi-layer query encoding strategy, enabling queries to align with document content at three levels: semantic, paragraph, and topic. The query vector is mapped to different subspaces through multi-head projection, forming:

$$H_q=\{h_q^{(1)},h_q^{(2)},...,h_q^{(K)}\} \tag{5}$$

The document slice vectors are also used to construct a corresponding multi-scale representation H_q, thereby achieving cross-scale matching during the retrieval stage. The matching score is obtained through weighted aggregation.

$$S_i=\sum_{k=1}^{K}\alpha_k sim(h_q^k,h_{d_i}^k) \tag{6}$$

The weight α_k is generated by an adaptive attention mechanism, enabling the retrieval to focus on semantic dimensions relevant to reasoning. Finally, candidate documents are selected based on normalized probabilities.

$$P(d_i|q)=\frac{exp(S_i)}{\sum_j exp(S_j)} \tag{7}$$

This ensures that the search results remain stable and structurally sensitive in long text scenarios.

2.3 Semantic Conflict Detection and Evidence Calibration

Long texts often contain heterogeneous content across paragraphs, topics, and even author sources. Therefore, conflict detection and semantic calibration mechanisms are needed to ensure the consistency of evidence information input into the generation stage. First, a semantic consistency matrix is constructed among all candidate evidence:

$$C_{ij}=g(h_d,h_{d_j}) \tag{8}$$

This is used to measure potential conflicts or support relationships between different segments. Then, graph regularization is used to penalize inconsistent evidence, causing the overall evidence set to tend towards semantic convergence.

$$L_{cons}=\sum_{ij}(1-C_{ij})\|z_i-z_j\|^2 \tag{9}$$

Based on this, the system constructs a global semantic center:

$$z_*=\frac{1}{M}\sum_{m=1}^{M}z_m \tag{10}$$

Used to represent a calibrated, unified semantic structure, ensuring cross-paragraph consistency of evidence at the final input generation end and reducing the propagation of error chains.

2.4 Evidence-Guided Conditional Generation

During the generation phase, to ensure that the reasoning process closely revolves around the retrieved evidence, the model constrains the generation distribution through conditional control and evidence attention mechanisms. Given the current generation state s_t and the integrated evidence set ε, the generation probability is obtained through evidence-aware attention:

$$a_t=softmax(\frac{s_tE^T}{\sqrt{d}}) \tag{11}$$

And use it for weighted aggregation of evidence representations:

$$e_t=\sum_t a_{t,i}E_i \tag{12}$$

Ultimately, the generated data is distributed within the language model and computed by fusing query semantics, historical generation states, and evidence representations:

$$p(y_t|y_{<t},q,\varepsilon)=softmax(W[s_t;e_t;h_q]) \tag{13}$$

This mechanism ensures that the output not only remains faithful to the content of the evidence but also maintains coherence in the logical chain, giving the reasoning process of long texts structured constraints and greater controllability.

3. Experimental Analysis

3.1 Dataset

The long-text reasoning dataset used in this study is derived from the HotpotQA Fullwiki Long-Context Version. Its main feature is that each sample consists of several interrelated Wikipedia documents. These documents are long and cover broad topics. This setting reflects the complexity of cross-paragraph and cross-document reasoning in real scenarios. The dataset contains multiple levels of textual structure, including introductions, background descriptions, factual statements, and related knowledge points. The model must integrate semantic clues from many sources during reasoning. This gives the dataset clear long-chain reasoning characteristics.

To meet the requirements of the retrieval-augmented generation framework, each sample includes a query, a set of candidate full documents, and a long sequence composed of several paragraphs. The information density inside the documents is uneven. Key evidence is often scattered across different sections or different documents. Some samples also contain conflicting or ambiguous content. The model must locate evidence across paragraphs, detect conflicts, and align semantic content. The complexity of the long-text structure helps evaluate the model's stability in noisy and heterogeneous textual environments.

The dataset also contains cross-topic and multi-type information, with both structured and unstructured content. It covers many domains such as history, technology, society, and culture. This allows a detailed examination of the retrieval module's ability to cover long documents and the generation module's capacity for structured reasoning under complex semantic dependencies. Through its large scale, multi-document composition, and long textual format, the dataset provides rich material and structural challenges for the long-text reasoning framework proposed in this study. It also supports the advancement of the model toward real-world knowledge reasoning applications.

3.2 Experimental Setup

This study uses Qwen-7B as the main backbone model for generation and reasoning. It serves as the conditional generator in the retrieval-augmented framework. The model is first tuned on a public long-text reasoning dataset with a lightweight instruction format. The tuning process adapts the model to the input style that combines a query and structured retrieved evidence. The model then generates natural language outputs that contain the reasoning chain and the final answer. The training procedure uses the AdamW optimizer. The learning rate is set to 1×10^{-5}. The batch size is 8. The maximum input length is set to 8k tokens to ensure that long text segments and multiple evidence contexts can be encoded at the same time. Gradient accumulation and mixed-precision computation are applied to reduce memory usage and speed up convergence. A length penalty and temperature control are used on the generation side to improve stability and diversity during long-chain reasoning.

Retrieval and model inference are performed in a unified experimental environment on the same hardware platform. This avoids additional interference caused by hardware differences. The retrieval component builds a semantic retrieval index based on vector embeddings. Document slices are encoded and used to construct the vector index offline. During inference, the system selects a fixed number of candidate evidence fragments from the index according to the query. These fragments are concatenated with the original query and fed to Qwen-7B for conditional generation. To ensure comparability across different experimental settings, the number of retrieved candidates, maximum generation length, and sampling hyperparameters such as top-k and top-p remain unchanged throughout all experiments. The

same random seed and data split strategy are used for all experiments. This reduces the impact of randomness on the analysis of long-text reasoning performance.

3.3 Experimental Results

To begin with, we conduct a comparative evaluation against several representative long-text reasoning methods, and the performance of all models across EM, F1-Score, Accuracy, and Faithfulness Score is summarized in Table 1.

Table 1. Comparative experimental results

Method	EM	F1-Score	Acc	Faithfulness Score
Self-rag [8]	32.4	45.8	51.2	58.3
Gnn-rag [9]	38.7	52.9	57.4	64.1
Omnieval [10]	41.5	56.2	60.3	67.8
Rageval [11]	44.1	59.8	63.7	70.5
Ours	52.9	68.4	72.1	81.7

As the methods become progressively stronger, the four evaluation metrics increase in a roughly monotonic manner, reflecting a consistent enhancement in long-text reasoning ability. This pattern suggests that improvements in retrieval quality, evidence organization, and generation control are all translated into steady gains in both accuracy-oriented and faithfulness-oriented indicators. This indicates that long-text reasoning relies heavily on retrieval quality, evidence integration, and generation consistency. The weakest method, Self-RAG, performs significantly worse on all four metrics. Its limitations on EM and F1 are especially clear. This reflects the difficulty of traditional retrieval-augmented approaches in capturing key semantic clues when facing long documents with cross-paragraph and multi-evidence dependencies. It also shows that shallow retrieval and simple concatenation of context cannot build an interpretable or stable reasoning chain. This often results in missing or misused evidence in the generated outputs.

When structured reasoning and graph-based representations are introduced, as in GNN-RAG, OmniEval, and RAGEval, performance improves in clear steps. The Faithfulness Score increases substantially. This shows that a stronger understanding of relationships among evidence can reduce hallucination and improve factual consistency. The simultaneous gains in F1 and Acc indicate that these methods retrieve key evidence more effectively and integrate cross-document semantic information more accurately during generation. The performance of RAGEval further demonstrates that enhanced long-chain reasoning capability helps maintain logical consistency and answer accuracy in scenarios with complex semantic spans.

In contrast, the method proposed in this study achieves the highest scores on all four core metrics. The improvements in EM and Faithfulness Score are particularly significant. This result shows that the hierarchical retrieval, conflict detection, and semantic calibration components in the framework enhance evidence localization and structural alignment across segments. By performing structured filtering and consistency compression on key paragraphs in long texts, the model can construct a clearer reasoning path. The final outputs are therefore more closely grounded in the actual evidence. In addition, the evidence constraint mechanism in the generation stage reduces semantic drift and improves both logical quality and credibility.

Overall, these experimental results confirm the suitability and advantages of the proposed retrieval-augmented generation framework for long-text reasoning. Unlike previous approaches that rely on local vector matching or single-structure encoding, this study strengthens the reasoning chain from three dimensions: semantic consistency, structured evidence integration, and generation-side constraints. This enables the model to obtain more complete key information, maintain cross-paragraph semantic coherence, and produce more interpretable and trustworthy reasoning outputs. These improvements are important for

advancing long-text reasoning in real applications such as legal review, scientific reading assistance, and policy understanding.

In addition, we investigate how the choice of learning rate influences the optimization behavior of the framework and its final reasoning performance; the variation of the four metrics under different learning rate settings is reported in Figure 2.

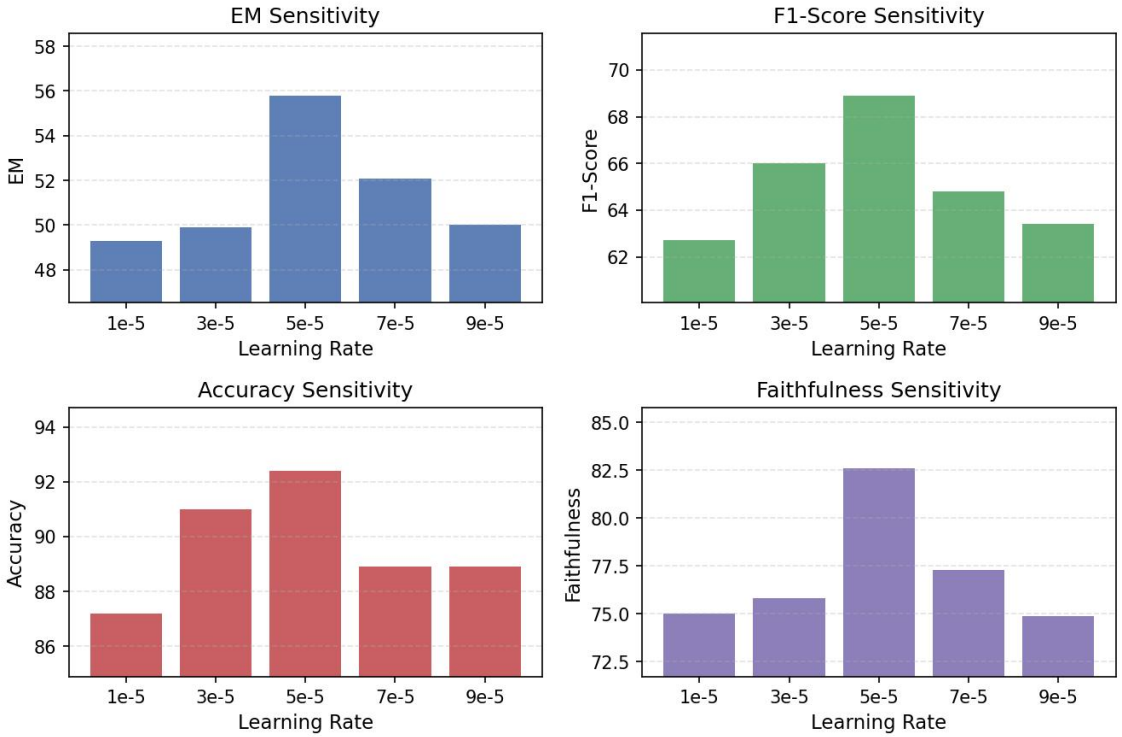

Figure 2. The impact of different learning rates on experimental results

Across different learning rate settings, the performance curves remain broadly stable, only exhibiting modest local oscillations rather than dramatic rises or collapses. This indicates that the proposed framework is not overly sensitive to the specific learning rate choice within a reasonable range, while still showing fine-grained variations that reflect the trade-off between convergence speed and semantic alignment. This indicates that the learning rate does not cause large performance swings in a retrieval-augmented generation framework. However, it still affects detailed behavior during evidence integration and semantic alignment. In particular, both EM and F1 decrease slightly when the learning rate is too low or too high. A moderate learning rate achieves the best performance. This aligns with the fact that both the generation module and the retrieval module participate in optimization for long-text tasks. A low learning rate fails to update semantic alignment structures effectively. A high learning rate tends to disrupt the continuity of evidence across paragraphs.

A closer look at the Accuracy and Faithfulness Score shows that Faithfulness exhibits larger variations. This means that the learning rate has a stronger influence on whether generated content remains faithful to retrieved evidence. Long-text reasoning depends on multi-evidence integration across paragraphs and documents. An improper learning rate can prevent effective calibration of conflicting evidence. This may lead to deviations or missing logic in the final output. An appropriate learning rate stabilizes updates in semantic calibration modules and evidence graph structures. This helps the model resolve conflicts more accurately and maintain content consistency.

Overall, the results show the important regulatory role of learning rate in long-text reasoning frameworks. A well-tuned learning rate improves answer accuracy and strengthens structural consistency and evidence faithfulness in the reasoning process. This produces outputs that align with the true semantic chain. For tasks involving large-scale cross-paragraph reasoning, an appropriate learning rate supports coordination across retrieval, alignment, calibration, and generation. This enhances the stability and robustness of long-text reasoning as a whole.

We further design a sensitivity study on the number of retrieved candidate documents, focusing on its effect on the Faithfulness Score. The relationship between retrieval breadth and evidence faithfulness is illustrated in Figure 3.

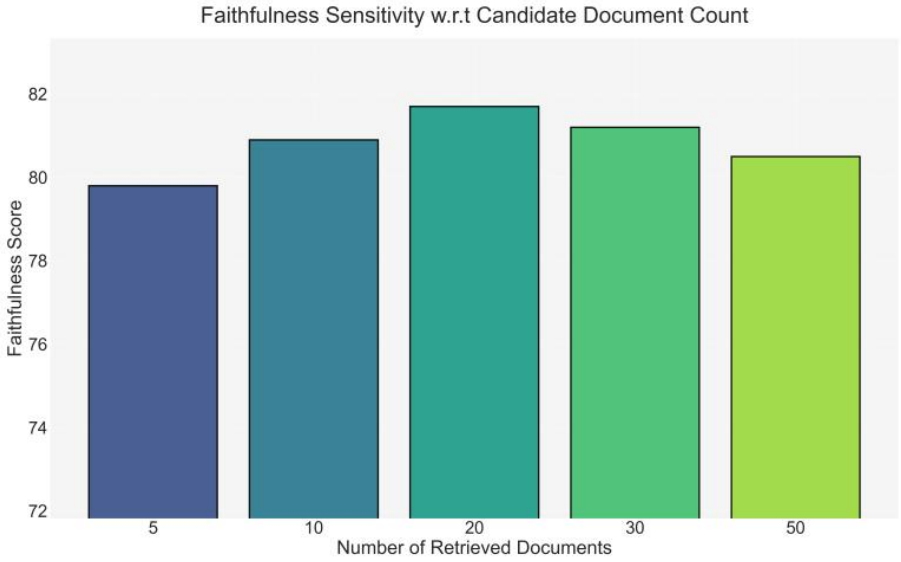

Figure 3. Sensitivity experiment of the number of candidate documents retrieved to the Faithfulness Score

When varying the number of retrieved candidate documents, the Faithfulness Score first increases noticeably and then tends to level off once a sufficient amount of evidence has been included. This pattern implies that expanding the evidence pool is beneficial up to a point, after which additional candidates contribute diminishing returns because most of the critical semantic cues have already been captured. This indicates that a moderate increase in retrieved evidence improves semantic consistency and factual faithfulness in long-text reasoning. When the number of candidates is too small, the model cannot access enough cross-paragraph information. This leads to incomplete evidence coverage in the reasoning chain. As a result, the generated content deviates from the true semantics. This is reflected by the low faithfulness score shown in the figure.

When the number of candidates increases to a medium level, the model can retrieve cross-document associations and form a more complete evidence structure during semantic encoding. At this point, the evidence conflict detection and semantic calibration modules operate more effectively. They integrate information from multiple sources into a unified semantic center. This substantially enhances the consistency between the generated reasoning content and the original evidence. This is the main reason why the Faithfulness Score reaches its highest value when the number of candidates is 20.

However, when the number of candidate documents continues to increase, performance shows a slight decline. This indicates that too many candidates introduce redundant or noisy evidence. This increases the burden on the semantic calibration module. When the evidence pool becomes too large, the model must search a wider space to identify useful segments. This may weaken some key semantic signals or dilute

them with irrelevant information. As a result, the faithfulness of the final output decreases slightly. This phenomenon confirms a typical challenge in retrieval-augmented frameworks for long-text tasks. Both insufficient information and information overload can harm reasoning quality.

Overall, the results demonstrate that controlling the number of retrieved candidates is essential for building a stable long-text reasoning chain. An appropriate number of documents allows the model to balance cross-paragraph evidence integration and semantic consistency. This maximizes the effectiveness of the semantic calibration module. The findings also highlight the importance of the hierarchical retrieval and semantic conflict correction mechanisms proposed in this study. These mechanisms help the model extract key evidence, filter noise, and produce reasoning outputs that are more faithful and more interpretable in complex long-text settings.

Moreover, we analyze how different context window length limits impact the model's ability to integrate cross-paragraph evidence and maintain semantic consistency. The corresponding trends in the evaluation metrics under various window sizes are presented in Figure 4.

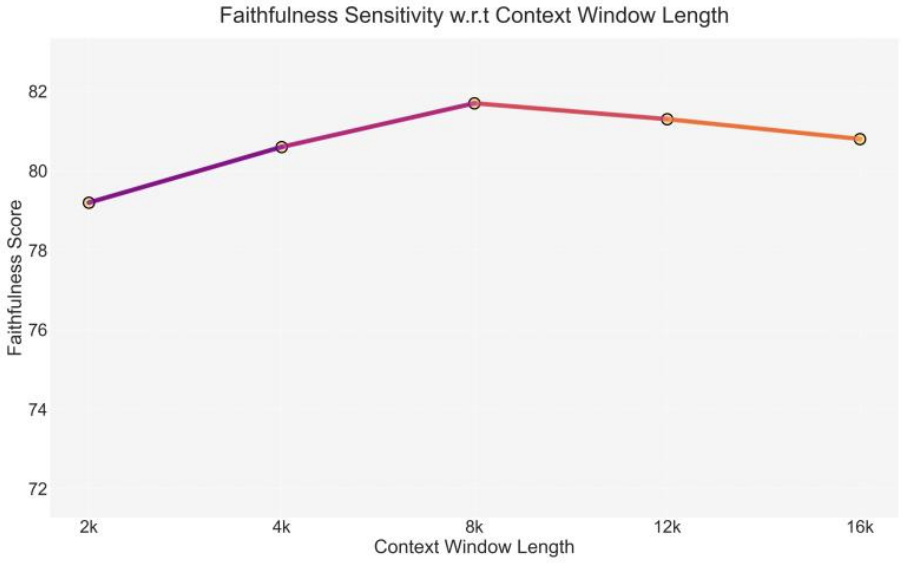

Figure 4. The impact of context window length limitation on experimental results

As the context window length grows, the Faithfulness Score initially benefits from the enlarged receptive field and then experiences a slight drop when the window becomes excessively long. This reveals that an intermediate window size provides the best balance between covering cross-paragraph dependencies and avoiding the distraction and noise introduced by overly large contexts. This trend indicates that the effective context range the model can process plays an important role in constructing the reasoning chain and maintaining semantic consistency. When the window is small, such as 2k or 4k, the model cannot receive enough cross-paragraph evidence. Important retrieved semantics are truncated. The relations within the reasoning chain cannot be fully preserved. This leads to lower faithfulness scores.

When the window length expands to a medium scale, such as 8k, the model can integrate more structured evidence in a single inference pass. This includes cross-document references, paragraph-level logical links, and local jump-style dependencies. Under these conditions, the semantic conflict detection and evidence calibration modules work in a richer information space. The generated content aligns more closely with the original evidence. As a result, the Faithfulness Score reaches its highest value. This shows that extending the context window enhances the model's ability to capture long reasoning chains.

However, when the window enlarges further to 12k or 16k, performance shows a slight decline. This suggests that large contexts introduce new challenges. Longer texts bring more noise evidence and dilute

the attention weight assigned to effective evidence. The model may shift away from key semantics during alignment. In addition, the generation module must handle more irrelevant information. This increases the burden on semantic calibration and reduces the faithfulness and coherence of reasoning. This indicates that larger context windows do not bring unlimited benefits.

Overall, the results confirm that context window length is a critical configuration factor in long-text reasoning frameworks. A moderate window provides the best balance between information coverage and noise suppression. This allows structured retrieval, conflict detection, and evidence calibration modules to work together and achieve the highest factual consistency. The trend also suggests that practical deployment of long-text reasoning models should consider task-specific needs when selecting context length rather than pursuing larger windows without limitation.

Finally, we examine the influence of training data scale on the overall long-text reasoning capability of the framework. The changes in EM, F1-Score, Accuracy, and Faithfulness Score as the proportion of training data increases from a small subset to the full dataset are depicted in Figure 5.

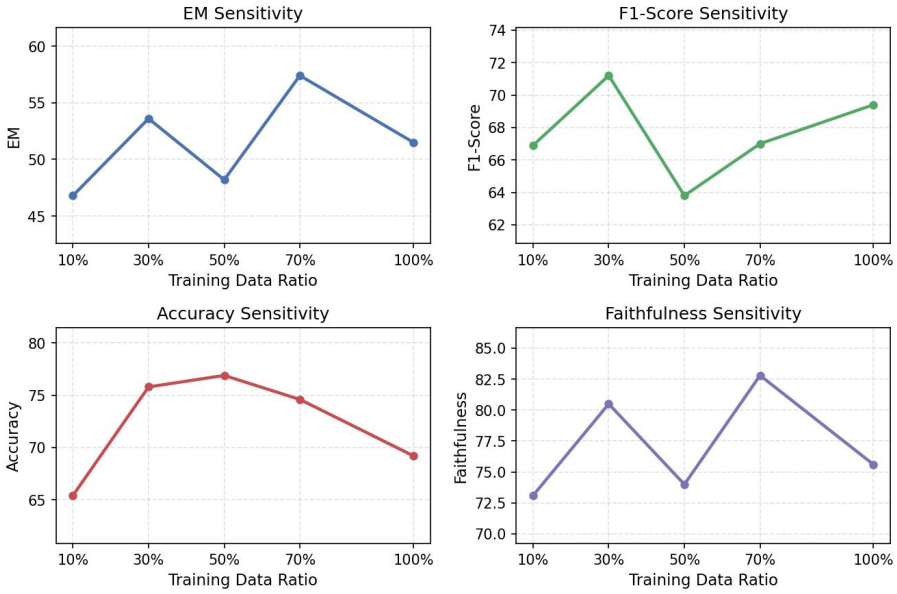

Figure 5. The impact of the proportion of training data on experimental results

When the proportion of training data is gradually increased, the model's long-text reasoning performance improves in a clear and consistent manner across all metrics. This demonstrates that larger training sets allow the framework to learn richer cross-document patterns, more robust evidence integration strategies, and more reliable reasoning behaviors for complex long-context scenarios. All four metrics show stable upward curves. This indicates that the retrieval-augmented generation framework proposed in this study has strong scalability under data expansion. The model learns richer semantic patterns, cross-paragraph dependencies, and evidence integration behaviors from larger datasets. When only 10 percent of the data is used, the model shows low EM, F1, Acc, and Faithfulness scores. This reflects that limited samples are not sufficient for learning the complex structural information required for long-chain reasoning.

When the training data increases to 50 percent, all metrics rise sharply. This pattern is characteristic of long-text tasks. The model must learn not only basic semantic matching but also paragraph-level logical structures, evidence combination strategies, and approaches for handling inconsistent information. Expanding the dataset strengthens the cooperation among the semantic calibration module, the conflict detection mechanism, and the structured retrieval component. The model can therefore construct more effective cross-paragraph reasoning chains and achieve significant improvements across multiple metrics.

When the training data approaches 100 percent, the performance improvement becomes less pronounced. This suggests that the model has reached a stable reasoning capability under large-scale training conditions. At this stage, additional data does not yield substantial gains. This reflects a saturation effect in structural reasoning ability for long-text models. At the same time, the stable increases in EM, F1, and Accuracy indicate that the model has developed strong cross-document semantic extraction and multi-evidence integration capabilities. It no longer relies on memorizing individual patterns but instead forms more robust reasoning strategies.

Overall, the results highlight the importance of training data scale for long-text reasoning models. Long-text tasks rely on complex cross-paragraph associations, and these associations must be gradually learned and structured through large quantities of samples. As the data scale grows, semantic consistency, factual faithfulness, and structural reasoning ability all improve. These findings confirm the effectiveness and scalability of the proposed framework in multi-evidence fusion and long-chain reasoning scenarios.

4. Conclusion

This study addresses key challenges in long-text reasoning, including insufficient retrieval, semantic drift, and inconsistent evidence. It proposes a retrieval-augmented generation framework that integrates hierarchical retrieval, semantic conflict detection, and evidence calibration. By introducing a structured evidence graph and cross-paragraph semantic alignment strategies, the model obtains more complete and stable semantic support in complex, multi-source, and cross-document environments. This significantly improves reasoning accuracy and consistency. The overall design optimizes the long-text reasoning process across three stages: evidence acquisition, semantic integration, and conditional generation. This allows the generated output to align more closely with the true semantic chain, reduces hallucination, and provides a new solution for tasks that involve multiple paragraphs, multiple documents, and strong logical dependencies.

The experimental results further confirm the reliability and advantages of the framework across several core metrics, including answer accuracy, reasoning stability, and factual faithfulness. Sensitivity experiments show how learning rate, context window size, retrieval candidate number, and data scale influence long-text reasoning performance. These experiments also demonstrate that the modular structure proposed in this study maintains robust performance under different training and inference conditions. This stability indicates that the framework is suitable not only for standard knowledge-based or factual reasoning tasks but also for more complex applications, such as legal analysis, scientific literature understanding, medical explanation reasoning, and accountability review.

From an application perspective, this study provides a practical technical foundation for long-text understanding systems across multiple domains. With the rapid growth of large language model applications in legal assistance, financial risk report analysis, policy interpretation, and technical document processing, the precision and interpretability of long-text reasoning have become critical factors for reliable systems. The retrieval-augmented framework proposed in this work improves the trustworthiness and controllability of the reasoning process through structured evidence control and semantic calibration. This offers strong foundational support for real-world system deployment.

Future research can explore dynamic structured retrieval, automatic modeling of cross-document graph topology, and enhanced control over reasoning chain generation. In real application scenarios, it will be necessary to incorporate domain knowledge bases, expert rule systems, or multimodal evidence to further strengthen reliable reasoning under complex contexts. The framework may also be extended to larger language models and longer context windows. Such extensions could reveal their potential for high-dimensional information integration, cross-domain knowledge transfer, and real-time interactive reasoning. These directions will help establish a stronger technical foundation for next-generation intelligent text analysis and decision support systems.

References

[1] D. Su, X. Li, J. Zhang, et al., "Read before generate! faithful long form question answering with machine reading", arXiv preprint arXiv:2203.00343, 2022.

[2] Z. Li, C. Li, M. Zhang, et al., "Retrieval augmented generation or long-context LLMs: A comprehensive study and hybrid approach", arXiv, 2024.

[3] H. T. Chen, F. Xu, S. Arora, et al., "Understanding retrieval augmentation for long-form question answering", arXiv preprint arXiv:2310.12150, 2023.

[4] Q. Leng, J. Portes, S. Havens, et al., "Long context RAG performance of large language models", arXiv preprint arXiv:2411.03538, 2024.

[5] Q. Zhao, R. Wang, Y. Cen, et al., "LongRAG: A dual-perspective retrieval-augmented generation paradigm for long-context question answering", arXiv preprint arXiv:2410.18050, 2024.

[6] Y. Lu, "Policy-Aware Enterprise Risk Assessment through LLM-Based Cross-Source Representation and Reasoning," 2024.

[7] C. S. Lee, "Long-Range Dependency Modeling and Decision Point Summarization for Large Language Models in Dialogue and Meeting Scenarios," 2024.

[8] A. Asai, Z. Wu, Y. Wang, et al., "Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection", 2024.

[9] C. Mavromatis and G. Karypis, "GNN-RAG: Graph Neural Retrieval for Large Language Model Reasoning", arXiv preprint arXiv:2405.20139, 2024.

[10] Y. Xue, "Enhancing Trustworthiness of Retrieval-Augmented Large Language Models via Confidence Calibration and Selective Rejection," 2024.

[11] H. Cui, "Task Graph Modeling and Constraint-Aware Scheduling for Reliable Large Language Model Agent Execution," 2024.