DuQUAD: A Dual-View Framework with Quality-Aware Evidence Pruning for Multi-Document Question Answering

Anonymous ACL submission

Abstract

Large Language Models with Retrieval-Augmented Generation perform well on knowledge-intensive question answering, but often fail to utilize relevant evidence in long-context multi-document settings due to positional bias, known as the Lost-in-the-Middle phenomenon. Existing mitigation strategies, including document reordering and attention steering, are fragile in noisy settings, where boundary bias and spurious relevance suppress mid-context evidence. We propose DuQUAD, a multi-document QA framework that mitigates positional bias via dual-view reasoning and quality-aware evidence pruning. It combines a Local Agent guided by a Structural Fusion Score with a Global Agent over the full context, enabling complementary local recovery and global coverage. Candidate evidence from both agents is filtered by explicit quality scoring and refined at the sentence level to suppress noise. DuQUAD consistently outperforms strong Lost-in-the-Middle mitigation baselines, including recent multi-agent and context optimization methods, achieving up to 13.8% improvement in answer accuracy and up to 9.4% improvement in golden document recall.

1 Introduction

Large Language Models (LLMs) have demonstrated strong performance on knowledge-intensive tasks (Achiam et al., 2023; Yang et al., 2025; Dubey et al., 2024), yet their inherently static knowledge leads to limitations like obsolescence and hallucination. Retrieval-Augmented Generation (RAG) mitigates these issues by grounding model outputs in retrieved documents (Lewis et al., 2020; Gao et al., 2023; Chuang et al., 2023). However, RAG suffers from a critical bottleneck in long-context settings known as the Lost-in-the-Middle (LiTM) effect (Liu et al., 2024). LLMs exhibit strong positional bias, disproportionately attending to documents at the input boundaries while underutilizing intermediate evidence. Consequently, even relevant documents placed in the middle often fail to contribute to reasoning, resulting in degraded multi-document QA performance.

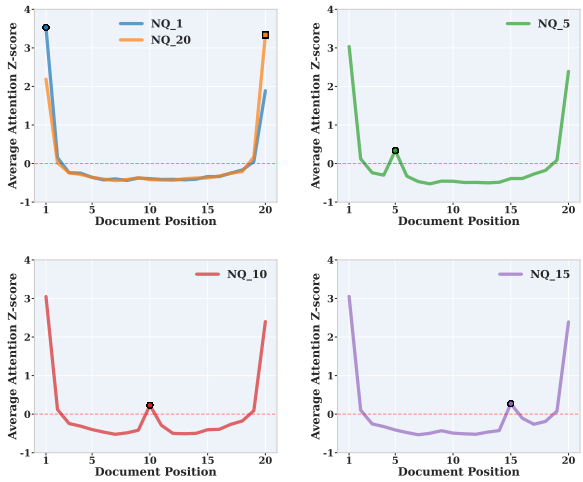

Figure 1: Attention z-score distribution across document positions on the Natural Questions (NQ) benchmark (NQ-1 to NQ-20, where NQ-n denotes instances in which the gold document appears at the n-th position).

Consistent with the well-known U-shaped positional bias (Hsieh et al., 2024), the model predominantly concentrates its attention on the beginning and end of the context. Distinct localized spikes emerge when gold documents appear in intermediate positions, forming a characteristic W-shaped pattern (Figure 1). However, these signals are often overshadowed by the overwhelming attention at the boundaries, causing the model to remain heavily distracted by the context start and end and fail to effectively attend to and utilize the crucial evidence located in the middle. To mitigate such positional bias, most existing approaches rely on document reordering or attention modification during training or inference (Hsieh et al., 2024; Jiang et al., 2024). However, relying on a weakly calibrated evidence signals such as semantic similarity or raw attention

renders them vulnerable to noise, where relevant evidence can be discarded or obscured by irrelevant documents.

Motivated by these localized spikes, we propose a framework that mitigates positional bias through dual-view reasoning and quality-aware pruning. We explicitly operationalize the W-shaped attention pattern as an attention contrast signal, integrating it with semantic and structural inference via a Structural Fusion Score (SFS) to prioritize gold documents. Our approach combines a Local Agent operating on prioritized documents with a Global Agent reasoning over the full context. Their outputs undergo DocScore-based filtering to prevent noise and echo chamber effects (Estornell and Liu, 2024), before a Judge Agent synthesizes the final answer.

We evaluate our framework on multi-document QA benchmarks covering controlled position and multi-hop reasoning scenarios. We compare our method against recent baselines, including LiTM mitigation strategies and multi-agent frameworks. Our approach demonstrates consistent performance improvements in these long and noisy multi-document environments. Our contributions are summarized as follows:

- We formalize the W-shaped attention pattern by adapting the edge-detection principle of the Laplacian operator (Marr and Hildreth, 1980) to effectively capture sharp local attention spikes amidst smooth positional bias.
- We propose the Structural Fusion Score (SFS), a unified document scoring function that integrates attention contrast signals, semantic relevance, and structural inference, enabling effective identification and re-ranking of documents for local evidence selection in multi-document QA settings.
- We introduce a Dual-View Agent Framework that mitigates positional bias by synergizing a global context view with a precise local view, integrated with quality-aware pruning to effectively filter noise and prevent hallucination.
- Extensive experiments across three multi-document QA benchmarks demonstrate that DuQUAD consistently improves answer accuracy by up to 13.8% and golden evidence recall by up to 9.4%, confirming its robustness to positional bias and noise in long-context settings.

2 Related Work

Retrieval-Augmented Generation (RAG) in Long-Context RAG has become a standard paradigm for grounding LLMs with external knowledge, addressing issues such as hallucination and knowledge staleness (Lewis et al., 2020). Prior work has explored reranking and evidence selection mechanisms to mitigate context contamination and superficial relevance in order to improve robustness (Gao et al., 2023; Chuang et al., 2023). However, in long-context multi-document settings, LLMs often exhibit strong positional bias, disproportionately attending to information near the beginning or end of the input while underutilizing evidence in the middle, a phenomenon known as Lost-in-the-Middle (LiTM) (Liu et al., 2024).

To alleviate LiTM, existing approaches have proposed prompt-level strategies such as document reordering (Jin et al., 2024) and context compression based on statistical metrics like perplexity (Jiang et al., 2024). These methods primarily focus on modifying input structure or reducing context length to improve usability in long-context settings.

Attention and Positional Bias in Language Models Attention distributions in Transformer-based models have long been studied as a means of analyzing model focus and interpretability (Vig, 2019). Recent analyses further show that attention patterns in long-context settings exhibit structured characteristics, including U-shaped positional bias and localized attention peaks, which may correlate with gold evidence (Hsieh et al., 2024). Motivated by these observations, several studies have explored leveraging attention-based signals for document importance estimation and evidence selection.

Multi-Agent Reasoning Multi-agent architectures have been proposed to enhance the robustness of RAG systems through iterative refinement, debate-style reasoning, or hierarchical query decomposition (Wu et al., 2024; Chang et al., 2025). While such frameworks expand evidence coverage and reasoning diversity, multi-stage pipelines remain vulnerable to error accumulation, particularly when early-stage evidence is noisy or weakly relevant (Mialon et al., 2023; Estornell and Liu, 2024; Wang et al., 2025).

3 Methodology

We address reasoning failures in long-context multi-document QA. Despite accurate retrieval, standard

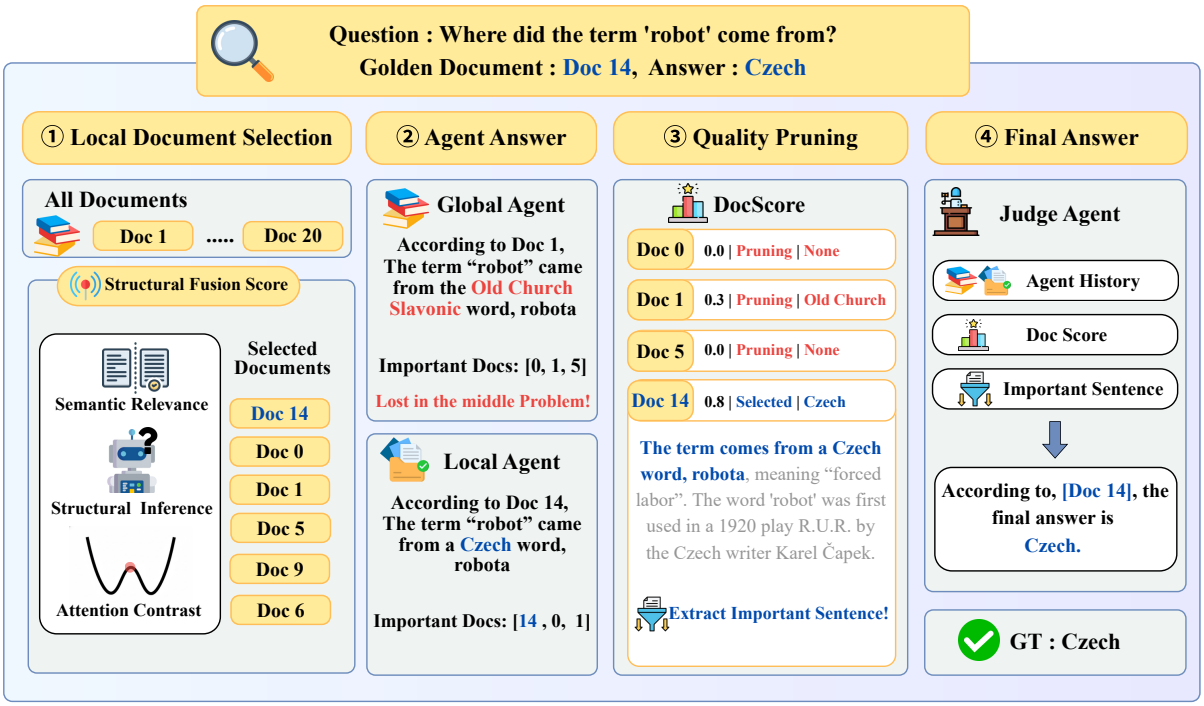

Figure 2: Overview of a dual-view multi-agent reasoning framework based on quality-aware evidence pruning. Given a question and a set of retrieved documents, the Structural Fusion Score (SFS) prioritizes a subset for the Local Agent, while the Global Agent operates on the full context. Selected documents are filtered via LLM-based DocScore, and key sentences are extracted. Finally, the Judge Agent aggregates reasoning histories and extracted evidence to produce the final answer.

RAG models suffer from positional bias and the LiTM effect, where mid-context evidence is suppressed by noise. To reliably isolate gold evidence, we propose a hierarchical framework that integrates document prioritization, dual-view agent collaboration, and quality-aware evidence pruning, as illustrated in the Figure 2.

3.1 Structural Fusion Score (SFS)

To identify gold documents under positional noise, we propose the Structural Fusion Score (SFS). While long-context attention typically exhibits a U-shaped bias, gold evidence forms a characteristic *W-shaped* pattern with sharp local spikes (Hsieh et al., 2024). SFS combines (1) semantic relevance, (2) structural inference, and (3) attention contrast. The first two components capture the alignment between the document and the query, while attention contrast explicitly emphasizes localized spikes that are robust to positional bias.

3.1.1 Semantic Relevance (Φ_{sim})

To capture the semantic alignment between a query and documents, we utilize the General Text Embeddings (GTE) model (Li et al., 2023)[1]. Given a query Q and a document D_i, we compute their dense vector representations $\mathbf{v}_Q$ and $\mathbf{v}_{D_i}$ independently. The semantic relevance is quantified as their cosine similarity:

$$\Phi_{\text{sim},i} = \frac{\mathbf{v}_Q^\top \mathbf{v}_{D_i}}{\|\mathbf{v}_Q\| \, \|\mathbf{v}_{D_i}\|} \tag{1}$$

3.1.2 Structural Inference (Φ_{ppl})

To capture structural inference signals, we evaluate the causal dependency of the query $Q = (q_1, \ldots, q_{|Q|})$ on a document D_i by measuring the likelihood of generating the query conditioned on the document context. This metric reflects intrinsic generative support beyond static embeddings.

$$\Phi_{\text{ppl},i} = -\frac{1}{|Q|} \sum_{t=1}^{|Q|} \log P(q_t \mid D_i, q_{<t}) \tag{2}$$

Here, $P(q_t \mid D_i, q_{<t})$ denotes the token-level generation probability assigned by the language model. Lower Φ_{ppl} values indicate stronger contextual support for the query, serving as a robust indicator of gold evidence.

[1]https://huggingface.co/Alibaba-NLP/gte-large-en-v1.5

3.1.3 Attention Contrast (Φ_{cont})

To isolate salient information peaks from positional noise, we compute a second-order local contrast over document-level attention scores. While absolute attention values in long contexts are heavily entangled with positional bias, sharp local deviations more reliably indicate gold evidence, revealing characteristic W-shaped patterns.

Here, a_i is computed as the average attention mass assigned to the tokens of document D_i from the final query position, aggregated over all layers and heads. In detail, we describe the attention score calculation in the Appendix I. Let a_i denote the aggregated attention score for document D_i, and let μ_A and σ_A be the mean and standard deviation over the retrieved set $A = \{D_1, \ldots, D_N\}$. We first standardize the scores:

$$z_i = \frac{a_i - \mu_A}{\sigma_A}.$$

We then define the structural contrast as a second-order difference:

$$\Phi_{\text{cont},i} = z_i - \frac{z_{i-1} + z_{i+1}}{2}.$$

This formulation is analogous to a discrete Laplacian operator (Gonzalez, 2009; Marr and Hildreth, 1980), suppressing smooth positional trends while emphasizing sharp local attention spikes.

3.1.4 Signal Fusion (Φ)

We integrate heterogeneous signals using a weighted softmax-based fusion. The final **Structural Fusion Score (SFS)** for document D_i is defined as:

$$\Phi_i = \alpha \mathcal{S}(\Phi_{\text{sim,i}}) + \beta \mathcal{S}(-\Phi_{\text{ppl,i}}) + \gamma \mathcal{S}(\Phi_{\text{cont,i}}) \tag{3}$$

where $\mathcal{S}(\cdot)$ denotes softmax applied over the retrieved documents. We negate Φ_{ppl} to ensure directional consistency.

Softmax is used to address scale heterogeneity and induce distribution sharpening, mapping disparate metrics into a unified probabilistic space while amplifying high-confidence evidence. We set $\alpha = 0.5$, $\beta = 1.0$, and $\gamma = 0.5$, placing higher weight on structural inference signals. Additional details are provided in Appendix D.

3.2 Dual-View Agent Collaboration

While SFS prioritizes relevant evidence, relying solely on a filtered subset risks losing global context and cross-document dependencies. To address this trade-off, we introduce a Dual-View Agent Collaboration framework that performs reasoning from two complementary views of the retrieved evidence.

The Global Agent reasons over the complete retrieved document set, preserving holistic context and cross-document relations, while the Local Agent focuses on a compact, noise-reduced subset re-ranked by SFS, enabling precise reasoning over highly prioritized evidence. This parallel design provides mutual robustness: the Local Agent recovers evidence suppressed by positional bias or noise, and the Global Agent compensates for information loss caused by aggressive filtering. Both agents independently generate answers with explicit citations, which are aggregated into a unified reasoning log, enabling robust and position-resilient multi-document reasoning.

3.3 Quality-Aware Evidence Pruning

While Dual-View reasoning maximizes recall, aggregating verbose reasoning logs and long-form documents can cause context overload, increasing the risk of hallucination. To address this issue, we introduce a two-stage quality-aware pruning strategy that explicitly separates evidence selection from quality verification.

Important Document Selection Each agent independently identifies a set of important documents based on its own reasoning process. This identification is directly derived from the agent's generated output. Specifically, each agent explicitly returns up to three document IDs as important documents, together with its reasoning results, following a predefined JSON output format.

Let $\mathcal{D}_{\text{G}}$ and $\mathcal{D}_{\text{L}}$ denote the sets of important documents selected by the Global agent and the Local agent, respectively. These sets are merged to form a candidate document pool as follows:

$$\mathcal{D}_{\text{imp}} = \mathcal{D}_{\text{G}} \cup \mathcal{D}_{\text{L}} \tag{4}$$

As a result, up to six candidate documents are obtained. This stage serves as a coarse, recall-oriented filtering step that broadly preserves answer-relevant evidence which might otherwise be missed due to positional bias or contextual noise.

LLM-Based Quality Scoring Each candidate document $D_i \in \mathcal{D}_{\text{imp}}$ is evaluated using an LLM-predicted DocScore $p_i \in [0, 1]$, obtained by prompting the LLM to assess whether the question

can be answered based solely on the given document. We retain only high-confidence documents using a threshold-based rule:

$$\mathcal{D}_{\text{keep}} = \{D_i \in \mathcal{D}_{\text{imp}} \mid p_i \geq \tau\}. \quad (5)$$

where the threshold $\tau = 0.4$ is fixed across all experiments. This value is empirically determined to balance golden document recall and noise reduction, as lower thresholds admit excessive noise while higher thresholds risk prematurely discarding gold evidence.

For each retained document, we generate a document-specific provisional answer and select the top two sentences as supporting evidence based on the length of word (or keyword) overlap between the answer and each sentence in the document. This sentence-level evidence compression reduces the computational burden on the Judge Agent while preserving the key information that directly contributes to final reasoning.

3.4 Final Decision via Judge Agent

The Judge Agent produces the final answer by jointly considering (i) the reasoning histories of the Global and Local Agents, (ii) document-level confidence scores (DocScore), and (iii) sentence-level evidence spans extracted from retained documents. Rather than relying on a single score or isolated evidence, the Judge integrates these heterogeneous signals to make a consolidated decision.

Formally, given the question Q, the reasoning traces $\mathcal{H}_{\text{G}}$ and $\mathcal{H}_{\text{L}}$, the sentence-level evidence set $\mathcal{S}$, and the associated DocScores $\mathbf{p}$ for documents in $\mathcal{D}_{\text{keep}}$, the final answer $\hat{y}$ is generated as:

$$\hat{y} = \text{JudgeAgent}(Q, \mathcal{H}_{\text{G}}, \mathcal{H}_{\text{L}}, \mathcal{S}, \mathbf{p}) \quad (6)$$

By aggregating complementary perspectives from multiple agents and evidence sources, the Judge Agent avoids over-reliance on any single document or signal, enabling robust and stable answer generation under positional bias and partial evidence loss.

4 Experiments

4.1 Experimental Setup

Backbone Model We use LLaMA-3.1-8B-Instruct (Dubey et al., 2024)[2] and Qwen2.5-7B-Instruct (Qwen et al., 2025)[3] as backbone models. Both support long-context reasoning and are applied consistently across all baselines to ensure fair comparison. Main results are reported for both models, while additional analyses use LLaMA-3.1-8B-Instruct for efficiency.

Datasets We evaluate our approach on three benchmarks. Natural Questions (NQ) (Kwiatkowski et al., 2019) is evaluated under the long-context protocol (Liu et al., 2024), where the position of gold documents is explicitly controlled ($k \in 1, 5, 10, 15, 20$) to analyze the LiTM effect. MuSiQue (Trivedi et al., 2022) and HotpotQA (Yang et al., 2018) are used as standard multi-hop QA benchmarks to assess reasoning over evidence distributed across multiple documents.

Evaluation Metrics We evaluate performance using accuracy, defined as whether the generated response contains the ground-truth answer string. This metric is commonly used for long-context QA with free-form outputs, focusing on answer presence rather than exact matching (Liu et al., 2024). For multi-hop reasoning tasks, we additionally report Exact Match (EM) for strict identity and F1 score for word overlap to assess generation precision.

4.2 Baselines

We compare our approach against baselines across three categories: (1) reordering methods including SBERT (Reimers and Gurevych, 2019)[4], BM25 (Crestani et al., 1998), and OpenAI embeddings (OpenAI, 2024)[5], (2) multi-agent reasoning frameworks such as Tree of Agents (ToA) (Yu et al., 2025) and MAIN-RAG (Chang et al., 2025), and (3) context optimization techniques like Found in the Middle (FiM) (Hsieh et al., 2024) and LongLLMLingua (Jiang et al., 2024). Implementation details are provided in Appendix A.

Main Results Table 1 reports accuracy on NQ, MuSiQue, and HotpotQA, adding EM and F1 scores for multi-hop tasks.

On NQ, the Vanilla LLM baseline exhibits strong sensitivity to document position. When the gold document appears in intermediate or trailing positions, accuracy decreases from 73.5% at NQ-1 to 64.0% at NQ-20, resulting in a degradation of 9.5%.

[2] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[3] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

[4] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

[5] https://platform.openai.com/docs/models/text-embedding-ada-002

Model	Method	NQ-1	NQ-5	NQ-10	NQ-15	NQ-20	MuSiQue			HotpotQA		
		Acc.	Acc.	Acc.	Acc.	Acc.	Acc.	EM	F1	Acc.	EM	F1
Llama 3.1 8B Instruct	Vanilla LLM	73.5	70.5	65.6	67.0	64.0	33.1	19.9	27.3	73.8	52.4	68.4
	SBERT	70.5	70.3	70.4	70.1	70.3	24.9	18.7	27.0	76.8	52.2	68.0
	BM25	70.4	69.9	69.8	69.2	69.7	33.6	17.7	26.5	75.3	50.8	67.8
	OpenAI	70.2	70.5	71.4	68.9	70.5	34.7	18.3	27.7	76.7	52.5	68.9
	ToA	58.0	53.5	56.1	51.0	52.2	21.5	14.0	22.2	57.3	44.2	58.0
	MAIN-RAG	62.5	64.4	63.1	62.3	61.4	19.0	16.2	23.7	60.5	46.6	61.6
	FiM	75.3	67.0	65.6	64.7	65.4	38.6	12.0	22.4	76.9	37	60.6
	LongLLMLingua	67.4	68.9	67.1	67.9	66.4	30.8	15.7	24.9	70.2	49.7	65.6
	DuQUAD	**77.3**	**76.0**	**75.2**	**76.1**	**76.7**	**41.1**	**20.6**	**30.6**	**80.2**	**52.6**	**70.2**
Qwen 2.5 7B Instruct	Vanilla LLM	68.1	67.4	65.9	66.4	66.9	32.2	17.7	26.2	79.2	53.5	66.8
	SBERT	66.5	66.4	66.1	66.6	66.2	29.6	21.3	29.0	79.2	57.5	71.1
	BM25	63.3	63.5	63.8	63.3	63.6	29.2	18.6	26.6	80.1	56.8	69.5
	OpenAI	65.4	65.4	65.9	65.0	65.2	30.2	20.0	27.5	80.3	57.2	71.1
	ToA	66.0	51.0	46.1	42.5	44.5	13.1	13.5	21.3	50.1	40.0	51.6
	MAIN-RAG	62.0	60.8	60.1	61.5	61.9	27.0	14.7	22.3	68.0	52.4	65.3
	FiM	73.6	61.0	59.0	58.5	58.7	32.4	12.6	21.3	74.5	36.2	54.2
	LongLLMLingua	63.8	64.9	65.2	64.7	65.1	25.6	15.7	24.9	67.6	51.5	61.9
	DuQUAD	**75.5**	**72.6**	**72.6**	**70.2**	**75.1**	**36.7**	**20.6**	**31.0**	**82.0**	**60.7**	**76.0**

Table 1: Table 1: Main experimental results comparing LLaMA 3.1 8B and Qwen2.5 7B. The highest score in each column is shown in **bold**, and the second-highest score is underlined. Model names are aligned to the left in the first column.

This trend clearly illustrates that the LiTM phenomenon substantially impairs long-context reasoning.

Reordering-based methods partially mitigate the performance degradation observed in the Vanilla baseline. Unlike the Vanilla LLM, which exhibits sharp accuracy drops depending on document position, reordering-based approaches using SBERT, BM25, and OpenAI embeddings maintain relatively stable performance, with accuracy variations limited to approximately 0.3 to 0.7% across positions. However, despite this improved stability, reordering-based methods remain vulnerable to contextual noise, and their overall performance consistently falls short of DuQUAD, which achieves an accuracy of 76.7% under the NQ-20 setting.

Context optimization approaches provide only limited performance improvements. For example, FiM achieves an accuracy of 75.3% at NQ-1 but exhibits a pronounced degradation to 64.7% at NQ-15. Similarly, multi-agent reasoning frameworks such as ToA and MAIN-RAG experience substantial performance drops under LiTM conditions. In particular, ToA degrades from 58.0% to 51.0% in accuracy, with even larger declines observed in EM and F1 scores. These results indicate that, in the absence of effective evidence quality pruning and explicit mitigation of positional bias, prior approaches alone remain vulnerable in multi-document QA settings.

In contrast, DuQUAD achieves state-of-the-art performance across all document positions. At one of the most challenging LiTM settings, NQ-15, DuQUAD improves accuracy by 7.2% over the reordering-based approach using OpenAI embeddings. It also outperforms the multi-agent baseline MAIN-RAG by 13.8% and exceeds context optimization methods such as LongLLMLingua by 8.2%. These gains are accompanied by consistent improvements in EM and F1 scores, indicating more precise answer grounding rather than superficial matching. Importantly, the same trends are observed when using both the LLaMA 3.1 8B and Qwen2.5 7B backbones, demonstrating that the effectiveness of DuQUAD is not tied to a specific model architecture.

DuQUAD also shows robust improvements on multi-hop benchmarks. On MuSiQue, DuQUAD surpasses FiM by 2.5% in accuracy while also achieving higher EM and F1 scores. On HotpotQA, DuQUAD improves accuracy by 3.3% over FiM, with corresponding gains in EM and F1. These results suggest that explicit noise control and quality-aware evidence selection remain effective for general multi-document reasoning beyond positional

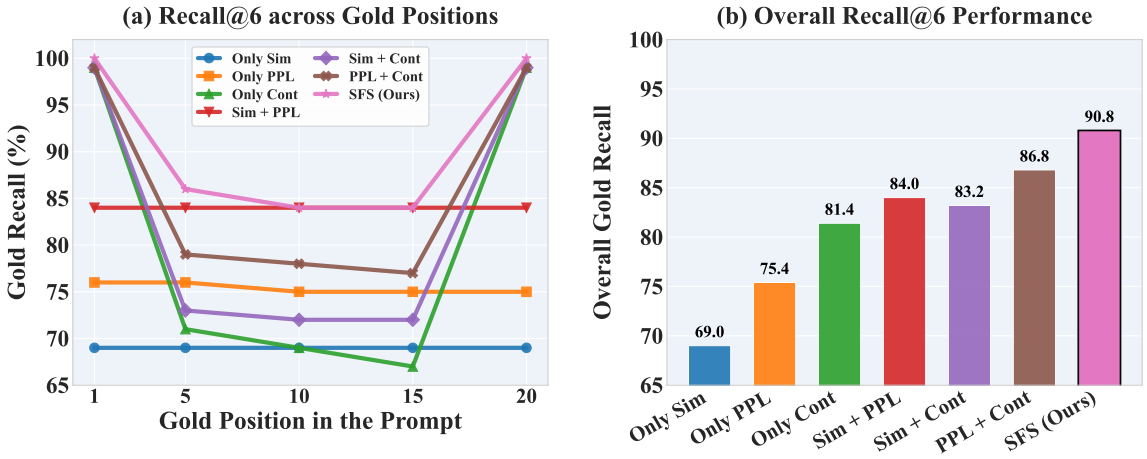

Figure 3: Impact of Signal Fusion on Robustness. (a) Recall of the gold document by gold position under the Top-6 setting. (b) Overall recall aggregated across all gold positions in (a).

Document Type	DS Mean	ID Avg.
Gold Doc	**0.647**	**85.7**
Non-Gold Doc	0.324	14.3

Table 2: Average DocScore (DS) and Important Document (ID) statistics on Natural Questions (NQ-1–20). DS Mean denotes the average DocScore assigned to documents of each type, while ID Avg. indicates the probability that a document of the corresponding type is selected as an Important Document.

Method	NQ-15	MuSiQue	HotpotQA
w/o Local Agent	69.8	38.1	79.1
w/o Quality Pruning	71.3	34.6	79.8
DuQUAD	**76.1**	**41.1**	**80.2**

Table 3: Effect of Local Agent and Quality Pruning. Ablation analysis across three datasets evaluating the impact on accuracy when omitting the Local Agent or Quality-Aware Pruning compared to the complete DuQUAD method.

bias.

Furthermore, paired bootstrap resampling (Berg-Kirkpatrick et al., 2012) confirms that the overall performance improvement is statistically significant, demonstrating the robustness of DuQUAD against random variations. Additional probabilistic statistical analyses of the main experimental results are provided in Appendix F.

Golden Evidence Recall We evaluate local document selection using golden evidence recall, defined as whether the gold document is included in the candidate set produced by the LocalAgent. As shown in Figure 3, methods relying on a single signal exhibit clear limitations. Attention Contrast (Cont) achieves the highest recall among single-signal approaches at 81.4%, surpassing Semantic Relevance (Rel) with 69.0% and Structural Inference (Inf) with 75.4%. In contrast, the proposed Structural Fusion Score (SFS) attains the highest overall recall of 90.8%, yielding a 4.0% improvement over the strongest multi-signal baseline that combines Inf and Cont, and a 9.4% gain over the best single-signal method. These results indicate that integrating complementary signals, including semantic relevance and generative relevance from Rel and Inf, together with localized attention peaks captured by Cont, enables more robust identification of gold evidence across different document positions.

Statistical Analysis of DocScore and Important Document Selection Table 2 reports aggregated statistics of DocScore and Important Document selection across all answer positions from NQ-1 to NQ-20, comparing gold and non-gold documents. Gold documents consistently receive substantially higher DocScore values than non-gold documents, with mean scores of 0.647 and 0.324 respectively, indicating that DocScore effectively captures intrinsic evidence quality. Moreover, gold documents are selected as Important Documents with high probability, averaging 85.7%, while non-gold documents are selected only rarely at 14.3%. This clear separation is consistently observed across all answer positions, suggesting that both DocScore assignment and Important Document selection are not driven by positional bias. Overall, these findings show that DocScore serves as a reliable quality signal that directly enables effective evidence prioritiza-

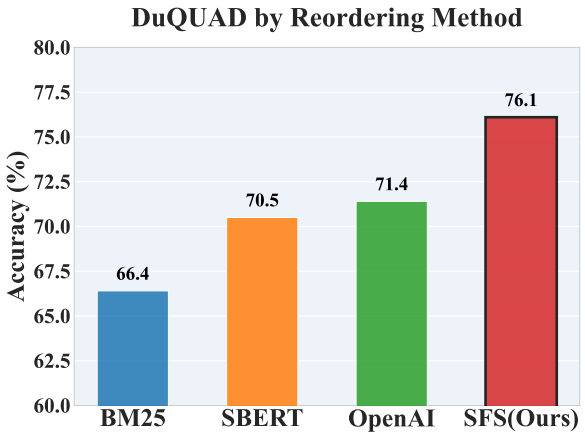

Figure 4: Impact of reordering mechanisms within DuQUAD on NQ-15. The proposed SFS is compared against baseline reordering methods to evaluate accuracy.

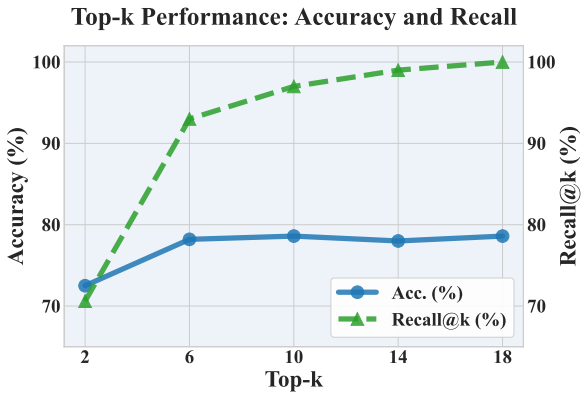

Figure 5: Impact of the number of documents provided to the local agent on Accuracy and Recall@k.

tion for quality-aware pruning in multi-document question answering.

Ablation Analysis Table 3 reports ablation results evaluating the contributions of the Local Agent and quality-aware evidence pruning. Removing either component leads to consistent performance degradation across all benchmarks. Specifically, removing the Local Agent reduces accuracy by 6.3% on Natural Questions, 3.0% on MuSiQue, and 1.1% on HotpotQA, indicating that global-context reasoning alone is insufficient under long-context positional bias. Disabling quality-aware evidence pruning causes even larger drops in challenging settings, with accuracy decreasing by 4.8% on NQ and 6.5% on MuSiQue, demonstrating that unfiltered low-quality evidence severely degrades downstream reasoning.

Figure 4 further analyzes document prioritization under the NQ-15 setting. When alternative reordering scores are used for local document selection, the highest accuracy among reordering-based methods is achieved by OpenAI embeddings at 71.4%, which remains substantially lower than the 76.1% attained by the proposed Structural Fusion Score (SFS). These ablation results confirm that the performance gains of DuQUAD arise from the synergy between SFS-based document prioritization, dual-view agent collaboration, and quality-aware evidence pruning.

Recall–Accuracy Trade-off Analysis Figure 5 illustrates the trade-off between answer accuracy and Recall@k as the number of retrieved documents k increases on Natural Questions. Recall@k increases monotonically, rising from approximately 70% at $k = 2$ to over 93% at $k = 6$ and saturating thereafter. In contrast, answer accuracy exhibits a clear elbow point around $k = 6$. Accuracy improves substantially from 72.5% at $k = 2$ to 78.2% at $k = 6$, but shows only marginal gains or slight fluctuations beyond this point despite continued improvements in recall. This behavior indicates that additional documents primarily introduce contextual noise that interferes with reasoning. Based on this observation, we select $k = 6$ as a balanced operating point that achieves high recall while effectively limiting noise, which motivates the need for selective document prioritization and quality-aware evidence pruning.

5 Conclusion

We investigated the limitations of long-context multi-document QA, specifically focusing on the 'Lost-in-the-Middle' phenomenon. To address these issues, we propose DuQUAD, a dual-view reasoning framework that utilizes Structural Fusion Score (SFS) prioritization and quality-aware pruning. The SFS integrates semantic relevance, structural inference, and attention contrast to robustly identify gold documents despite positional noise. DuQUAD effectively preserves relevant evidence while suppressing noise, achieving SOTA performance with an accuracy improvement of up to 13.8% over baselines on NQ-15. Furthermore, SFS significantly refined local evidence selection, increasing gold evidence recall to 90.8%—a 9.4% gain over single-signal methods. These results demonstrate that DuQUAD is an effective framework for robust reasoning in long-context multi-document QA.

6 Limitations

While DuQUAD effectively mitigates positional bias in long-context multi-document QA, it also has several limitations. First, DuQUAD employs an explicit evidence pruning mechanism to suppress noisy or low-confidence documents. Although this design improves robustness against positional noise, it may reduce contextual coverage when relevant evidence is weakly expressed or distributed across multiple documents, particularly for complex multi-hop questions that require integrating subtle cues from several sources. Second, DuQUAD relies on the availability of answer-bearing documents within the initial retrieved set. Similar to other retrieval-augmented frameworks, its performance is bounded by retrieval quality: if relevant documents are not retrieved, the framework cannot recover missing evidence through internal reasoning alone. These limitations reflect inherent trade-offs in multi-document QA systems that balance contextual coverage, noise suppression, and reasoning stability.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in nlp. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 995–1005.

Chia-Yuan Chang, Zhimeng Jiang, Vineeth Rakesh, Menghai Pan, Chin-Chia Michael Yeh, Guanchu Wang, Mingzhi Hu, Zhichao Xu, Yan Zheng, Mahashweta Das, and 1 others. 2025. Main-rag: Multi-agent filtering retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2607–2622.

Yung-Sung Chuang, Wei Fang, Shang-Wen Li, Wen-tau Yih, and James Glass. 2023. Expand, rerank, and retrieve: Query reranking for open-domain question answering. *arXiv preprint arXiv:2305.17080*.

Fabio Crestani, Mounia Lalmas, Cornelis J Van Rijsbergen, and Iain Campbell. 1998. "is this document relevant?... probably" a survey of probabilistic models in information retrieval. *ACM Computing Surveys (CSUR)*, 30(4):528–552.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Andrew Estornell and Yang Liu. 2024. Multi-llm debate: Framework, principals, and interventions. *Advances in Neural Information Processing Systems*, 37:28938–28964.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1).

Rafael C Gonzalez. 2009. *Digital image processing*. Pearson education india.

Cheng-Yu Hsieh, Yung-Sung Chuang, Chun-Liang Li, Zifeng Wang, Long Le, Abhishek Kumar, James Glass, Alexander Ratner, Chen-Yu Lee, Ranjay Krishna, and 1 others. 2024. Found in the middle: Calibrating positional attention bias improves long context utilization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14982–14995.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2024. Longllmlingua: Accelerating and enhancing llms in long context scenarios via prompt compression. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1658–1677.

Bowen Jin, Jinsung Yoon, Jiawei Han, and Sercan O Arik. 2024. Long-context llms meet rag: Overcoming challenges for long inputs in rag. *arXiv preprint arXiv:2410.05983*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, and 1 others. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy

Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

David Marr and Ellen Hildreth. 1980. Theory of edge detection. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 207(1167):187–217.

Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, and 1 others. 2023. Augmented language models: a survey. *arXiv preprint arXiv:2302.07842*.

OpenAI. 2024. Text embedding 3 models. https://platform.openai.com/docs/models/text-embedding-3-large. Accessed: 2025-12-25.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. musique: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Jesse Vig. 2019. Visualizing attention in transformer-based language representation models. *arXiv preprint arXiv:1904.02679*.

Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan O Arik. 2025. Astute rag: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30553–30571.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, and 1 others. 2024. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First Conference on Language Modeling*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2369–2380.

Song Yu, Xiaofei Xu, Ke Deng, Li Li, and Lin Tian. 2025. Tree of agents: Improving long-context capabilities of large language models through multi-perspective reasoning. *arXiv preprint arXiv:2509.06436*.

A Baseline Implementation Details

(0) Vanilla LLM. Constructs the input context by directly concatenating one ground-truth document with 19 distractors without additional post-processing.

(1) Reordering. (i) SBERT Reordering (Reimers and Gurevych, 2019) A dense retrieval baseline using Sentence-BERT. Documents are rearranged in descending order of cosine similarity between the query and document embeddings. (ii) BM25 Reordering (Crestani et al., 1998) A sparse retrieval baseline utilizing lexical overlap. Documents are re-ranked based on BM25 scores to prioritize exact term matches. (iii) OpenAI Reordering (OpenAI, 2024) Reorders documents based on semantic similarity, leveraging the generalization of proprietary embeddings.

(2) Multi-Agent Reasoning. (i) Tree of Agents (ToA) (Yu et al., 2025) Organizes agents in a hierarchical tree topology. A root agent decomposes tasks while leaf agents explore distinct paths, aggregating insights to mitigate information loss in long contexts. (ii) MAIN-RAG (Chang et al., 2025) A collaborative architecture with specialized agents. It employs adaptive filtering to dynamically set inclusion thresholds, retaining high-confidence evidence while discarding noise.

(3) Context Optimization. (i) Found in the Middle (FiM) (Hsieh et al., 2024) A position-aware strategy countering the "Lost-in-the-Middle" phenomenon. It applies a U-shaped reordering, placing the most relevant documents at the beginning and end of the context. (ii) LongLLMLingua (Jiang et al., 2024) Focuses on prompt compression. Following the original implementation settings, it utilizes `Llama-2-7b-chat-hf`[6] to estimate token perplexity, discarding low-importance tokens to compress the prompt while preserving key information.

[6] https://huggingface.co/meta-llama/Llama-2-7b-chat-hf

Experiment	Dataset	Samples	Position
W-shape visualization	NQ	500	Figure 1
Main Experiment	NQ	5,000	Table 1
	MuSiQue	1,000	
	HotpotQA	1,000	
Gold Recall Analysis	NQ	500	Figure 3
DocScore & Important Document Statistical Analysis	NQ	5,000	Table 2
Local Agent & Quality Pruning Ablation	NQ-15	1,000	Table 3
	MuSiQue	1,000	
	HotpotQA	1,000	
Reordering Method Ablation	NQ-15	1,000	Figure 4
Top-k Performance	NQ-15	1,000	Figure 5

Table 4: Summary of experimental settings and datasets.

Threshold (τ)	Recall (%)
0.2	83.57
0.3	80.70
0.4	**80.70**
0.5	79.38
0.6	79.02
0.7	75.06
0.8	74.70

Table 5: Sensitivity of answer-document recall to the DocScore threshold τ (operating range). We report $\tau \in [0.2, 0.8]$ as extremely low thresholds behave similarly (near-pass-through), while overly high thresholds enter a degenerate pruning regime.

Φ_{sim}	Φ_{ppl}	Φ_{cont}	Avg. Recall (%)
0.5	**1.0**	**0.5**	**91.40**
0.3	0.5	0.3	91.20
0.3	0.7	0.3	91.20
0.3	0.7	1.0	91.20
0.3	0.5	0.7	91.00
0.3	0.5	1.0	91.00
0.3	0.7	0.7	91.00
0.3	1.0	0.3	91.00
0.5	0.7	1.0	91.00
0.7	0.7	0.5	91.00

Table 6: Grid search results for Structural Fusion Score (SFS) weights. Average golden evidence recall is reported across all evaluated answer positions. The selected configuration $(\Phi_{\text{sim}}, \Phi_{\text{ppl}}, \Phi_{\text{cont}}) = (0.5, 1.0, 0.5)$ achieves the highest overall recall.

B Experimental Settings

This appendix describes the experimental configurations used for the main results, ablation studies, and analytical experiments presented in the paper.The datasets and the number of samples vary depending on the experimental purpose. All results reported in this paper follow the settings summarized in Table 4.

C DocScore Threshold Sensitivity Analysis

Table 5 shows the recall of gold documents under varying DocScore thresholds. Recall decreases monotonically as τ increases, reflecting the expected trade-off between pruning aggressiveness and evidence coverage. Importantly, recall remains unchanged between $\tau = 0.3$ and $\tau = 0.4$ (80.70%), indicating that raising the threshold within this range does not discard additional gold documents. Beyond $\tau = 0.4$, recall consistently declines, suggesting that stricter pruning begins to remove gold evidence. We therefore set $\tau = 0.4$ as the highest threshold that preserves recall while still enabling meaningful pruning of low-confidence documents. These results are obtained on the NQ setting with 1,000 examples.

D Ablation on SFS Weight Selection

We determine the weighting coefficients of the Structural Fusion Score (SFS) via a grid search over $(\Phi_{\text{sim}}, \Phi_{\text{ppl}}, \Phi_{\text{cont}})$ conducted on the Natural Questions (NQ) dataset. Each configuration is evaluated using the average golden evidence recall aggregated across all answer positions. As shown in Table 6, multiple weight combinations achieve comparable recall values, indicating that SFS is robust to moderate variations in weighting. Among them, the configuration $(0.5, 1.0, 0.5)$ attains the highest overall average recall (91.4%) and is therefore adopted in all experiments. This choice places greater emphasis on structural inference signals while retaining complementary contributions from semantic relevance and attention contrast. These results are obtained on the NQ setting with 1,000 examples.

E Reproducibility

To ensure reproducibility, we release the exact prompts used for all LLM-based components. Reasoning agent prompts, Judge prompts, and score estimation prompts are provided in Figures 6, 7, and 8, respectively, without modification. All experiments were conducted on a computing system equipped with NVIDIA L40S GPUs, an AMD EPYC 9654 CPU (96 cores), and 1 TB of system memory. We fix all decoding parameters across experiments and use deterministic decoding with a temperature of 0.0 unless otherwise specified, enabling faithful replication of our results.

All baseline experiments were conducted under the same generation setting with the temperature fixed to 0.0 to ensure reproducibility and fair comparison. The basic prompt templates (e.g., Vanilla LLM and reordering-based methods) strictly follow the configurations used in the Lost in the Middle paper [7]. For baselines requiring additional prompt control or specialized templates, we prioritized the officially released GitHub implementations provided by the original authors. When an official implementation was not available, we faithfully reimplemented the prompts and inference procedures as described in the corresponding papers.

Structural inference and attention score computation are performed using the same backbone model as the main reasoning process. This avoids introducing separate scoring models and helps mitigate bias arising from reliance on specific model architectures.

F Statistical Significance Analysis

We evaluate the statistical significance of the performance differences between DuQUAD and each baseline using paired bootstrap resampling. For each dataset, we repeatedly resample the evaluation set with replacement (10,000 iterations) and compute the containment rate difference between DuQUAD and the corresponding baseline. The reported difference corresponds to the mean improvement, and the 95% confidence interval (CI) is obtained from the 2.5th and 97.5th percentiles of the bootstrap distribution.

Statistical significance is assessed using two-sided bootstrap hypothesis testing under the null hypothesis that the expected difference is zero. P-values are computed as the proportion of bootstrap samples whose sign contradicts the observed mean difference. We report significance levels using standard conventions: *** for $p < 0.001$, ** for $p < 0.01$, and * for $p < 0.05$.

The complete statistical results are summarized in Table 7. Across all datasets, DuQUAD consistently outperforms strong retrieval, reordering, compression, and multi-agent baselines with statistically significant margins. Notably, the improvements remain significant even against competitive methods such as OpenAI reordering, FiM, and LLMLingua, indicating that the observed gains cannot be attributed to random variation.

G Cost Analysis

Table 8 reports the computational cost of each method normalized to the Vanilla baseline. We measure cost using two complementary indicators: the total number of LLM calls and the total number of processed tokens. LLM calls reflect the number of model invocations in each pipeline, while token counts capture the effective computation incurred by long-context processing.

All values are averaged over the same evaluation instances and normalized with respect to the Vanilla setting. Ratios indicate relative overhead compared to Vanilla, with higher values corresponding to increased computational cost.

As shown in Table 8, DuQUAD requires moderately more LLM calls than single-pass baselines ($4.1\times$) due to its dual-view reasoning and evidence scoring stages. However, its token consumption remains comparable to lightweight reordering and compression-based methods (approximately $2.1\times$), and is substantially lower than multi-agent frameworks such as MAIN-RAG and ToA. This demonstrates that DuQUAD achieves strong performance gains while avoiding the extreme computational overhead typically associated with multi-agent reasoning.

H Boundary Handling for Attention Contrast Computation

When computing attention contrast, special handling is required at document boundary positions where symmetric neighbors do not exist. Instead of constructing a second-order central difference operator that assumes symmetric neighbors through padding or reflection, we adopt a one-sided finite difference scheme at the document boundaries. Specifically, for the first and last documents, at-

[7] https://github.com/nelson-liu/lost-in-the-middle

tention contrast is computed using the difference between the document and its nearest neighbor. This design reduces artificial smoothing effects introduced by padding-based boundary conditions while more faithfully preserving relative attention salience at the document edges. The one-sided difference scheme is widely used across numerical analysis and signal processing, and in our experiments, it exhibits consistent and interpretable behavior for document-level contrast estimation.

I Document-Level Attention Scoring

This section briefly describes how we compute document-level attention scores, which are used as one component of the structural fusion score.

Input sequence construction. Given a question Q and a set of retrieved documents $\{D_1, \ldots, D_K\}$, we follow the Lost-in-the-Middle setting and construct the input sequence as:

$$[Q; D_1; D_2; \ldots; D_K; Q].$$

Each document D_i is preceded by a special delimiter token (e.g., `[DOC]`), and the document text is used as provided by benchmark datasets. The token span of document D_i in the input sequence is recorded as $\text{span}(D_i) = [s_i, e_i)$.

Document-level attention score. We use self-attention from all transformer layers, averaged across heads and layers, and focus on the attention distribution corresponding to the final query token. The attention score for document D_i is defined as the mean attention mass assigned to the tokens belonging to its document span:

$$\text{AttnScore}(D_i) = \frac{1}{|\text{span}(D_i)|} \sum_{j \in \text{span}(D_i)} r[j].$$

```
"Write a high-quality answer for the given question using only "
"the provided search results (some of which might be irrelevant)."
"Use ONLY the provided documents. Cite DOC ids inline like [DOC i] with quoted snippets. "
"At the end, output a JSON object with this exact format:\n"
'{"important_docs": [<doc_id1>, <doc_id2>, <doc_id3>]}\n'
"Include the top 3 most important document IDs."

f"--- Context Documents ---\n{ctx_block}\n\n"
f"Question: {question}\n“
```

Figure 6: Prompt of agents

```
"You are a judge. Use ONLY the provided evidence snippets and debate log. "
"The DocScore section shows reliability scores (p=0.0-1.0). Higher scores are more reliable. "
"Write a high-quality answer for the given question using only "
"the provided search results (some of which might be irrelevant)."

f"--- DocScore (Document Reliability Scores) ---\n{docscore_block}\n\n"
f"--- Important Evidence Snippets ---\n{evidence_block}\n\n"
f"{docscore_section}"
f"--- Debate Log ---\n{debate_log}\n\n"
f"Question: {question}\n"
```

Figure 7: Prompt of judge agent

```
"For each doc, estimate how likely it answers the question. "
f"IMPORTANT: You MUST evaluate ALL {len(selected_doc_ids)} documents: {selected_doc_ids}. "
"Output a JSON object with this exact format:\n"
"{\n"
' "scores": [\n'
'  {"doc_id": <id>, "p": <0.0-1.0>, "answer": "<best answer span/phrase>"},\n'
'  {"doc_id": <id>, "p": <0.0-1.0>, "answer": "<best answer span/phrase>"}\n'
" ]\n"
"}\n"
f"Make sure to include scores for ALL document IDs: {selected_doc_ids}\n\n"
```

Figure 8: Prompt of Score agent

Dataset (DuQUAD Acc.)	Baseline	Accuracy(%)	Diff (95% CI)	P-value
NQ-15 (76.1)	Vanilla LLM	67.0	+9.1 (6.4 – 11.8)	< 0.001***
	SBERT	70.1	+6.0 (3.2 – 8.8)	< 0.001***
	BM25	69.2	+6.9 (4.1 – 9.6)	< 0.001***
	OpenAI	69.9	+6.2 (3.4 – 9.0)	< 0.001***
	ToA	51.0	+25.1 (21.8 – 28.4)	< 0.001***
	MAIN-RAG	62.3	+13.8 (10.9 – 16.7)	< 0.001***
	LongLLMLingua	67.9	+8.2 (5.3 – 11.1)	< 0.001***
	FiM	64.7	+11.4 (8.6 – 14.2)	< 0.001***
Musique (41.1)	Vanilla LLM	33.1	+8.0 (5.3 – 10.7)	< 0.001***
	SBERT	24.9	+16.2 (13.5 – 19.0)	< 0.001***
	BM25	33.6	+7.5 (5.0 – 10.0)	< 0.001***
	OpenAI	34.7	+6.4 (3.8 – 9.0)	< 0.001***
	ToA	21.5	+19.6 (16.6 – 22.6)	< 0.001***
	MAIN-RAG	19.0	+22.1 (19.2 – 25.0)	< 0.001***
	LongLLMLingua	30.8	+10.3 (7.5 – 13.2)	< 0.001***
	FiM	38.6	+2.5 (-0.4 – 5.4)	0.094
HotpotQA (80.2)	Vanilla LLM	73.8	+6.4 (3.9 – 8.9)	< 0.001***
	SBERT	76.8	+3.4 (1.0 – 5.8)	0.007**
	BM25	75.3	+4.9 (2.5 – 7.4)	< 0.001***
	OpenAI	76.7	+3.5 (1.0 – 6.0)	0.008**
	ToA	57.3	+22.9 (19.9 – 26.1)	< 0.001***
	MAIN-RAG	60.5	+19.7 (16.6 – 22.7)	< 0.001***
	LongLLMLingua	70.2	+10.0 (7.2 – 12.8)	< 0.001***
	FiM	76.9	+3.3 (0.7 – 5.8)	0.014*

Table 7: Performance comparison of Accuracy (%). Values in parentheses represent the accuracy of DuQUAD. Mean differences and 95% confidence intervals (CI) are derived from paired bootstrap resampling ($N = 10,000$). Asterisks indicate statistical significance: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Method	LLM Calls		Tokens	
	Value	Ratio	Value	Ratio
Vanilla LLM	1.00	1.00×	2864	1.00×
OpenAI	2.00	2.00×	5920	2.07×
FiM	2.00	2.00×	5437	1.90×
LongLLMLingua	2.00	2.00×	5822	2.03×
MAIM-RAG	41.0	41.0×	16406	5.73×
ToA	26.5	26.5×	23684	8.27×
DuQUAD	4.10	4.10×	5914	2.07×

Table 8: Cost Comparison (Normalized to Vanilla)

Algorithm 1 DuQUAD Pipeline for a Single Question

Input: question Q, documents $\{\mathcal{D}_i\}_{i=1}^{K}$, local top-$m$, threshold τ
Output: final answer $\hat{y}$

(1) Local selection via Structural Fusion Score (SFS)
$a_i \leftarrow$ aggregated attention score of $\mathcal{D}_i$
$\Phi_{\text{sim},i} \leftarrow$ semantic relevance score of $\mathcal{D}_i$
$\Phi_{\text{ppl},i} \leftarrow$ structural inference score of $\mathcal{D}_i$

$a_i \leftarrow$ aggregated attention score of document D_i

$$z_i \leftarrow \frac{a_i - \mu_A}{\sigma_A}$$ ▷ z-score normalization (μ_A, σ_A: mean/std of $\{a_i\}_{i=1}^{K}$)

$$\Phi_{\text{cont},i} \leftarrow z_i - \frac{z_{i-1} + z_{i+1}}{2}$$ ▷ attention contrast ($i = 2, \ldots, K-1$)

$$\Phi_i \leftarrow \alpha\,\text{softmax}(\Phi_{\text{sim}})_i + \beta\,\text{softmax}(-\Phi_{\text{ppl}})_i + \gamma\,\text{softmax}(\Phi_{\text{cont}})_i$$ ▷ structural fusion score

$\mathcal{I}_{\text{local}} \leftarrow \arg\text{top}_m\,\Phi_i$ ▷ local document selection
$\mathcal{D}_{\text{local}} \leftarrow \{\mathcal{D}_i \mid i \in \mathcal{I}_{\text{local}}\}$

(2) Dual-view Agent Collaboration
$(\mathcal{H}_G, \mathcal{D}_G) \leftarrow \text{GlobalAgent}(Q, \{\mathcal{D}_i\}_{i=1}^{K})$ ▷ global reasoning + important docs
$(\mathcal{H}_L, \mathcal{D}_L) \leftarrow \text{LocalAgent}(Q, \mathcal{D}_{\text{local}})$ ▷ local reasoning + important docs
$\mathcal{D}_{\text{imp}} \leftarrow \mathcal{D}_G \cup \mathcal{D}_L$ ▷ merge important docs

(3) DocScore-based Quality Gating
$(p_i, \tilde{a}_i) \leftarrow \text{ScoreAgent}(Q, \mathcal{D}_i), \quad \forall \mathcal{D}_i \in \mathcal{D}_{\text{imp}}$ ▷ document reliability + answer span
$\mathcal{D}_{\text{keep}} \leftarrow \{\mathcal{D}_i \in \mathcal{D}_{\text{imp}} \mid p_i \geq \tau\}$ ▷ DocScore-based pruning

(4) Evidence Snippet Extraction
$\mathcal{S}_i \leftarrow \text{Top2}_j\,\text{overlap}(u_{i,j}, \tilde{a}_i), \quad \forall \mathcal{D}_i \in \mathcal{D}_{\text{keep}}$ ▷ select evidence sentences
$\mathcal{S} \leftarrow \bigcup_{\mathcal{D}_i \in \mathcal{D}_{\text{keep}}} \mathcal{S}_i$ ▷ aggregate evidence

(5) Final Decision
$\hat{y} \leftarrow \text{JudgeAgent}(Q, \mathcal{H}_G, \mathcal{H}_L, \mathcal{S}, \{p_i\}_{\mathcal{D}_i \in \mathcal{D}_{\text{keep}}})$ ▷ final answer generation

return $\hat{y}$