

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

Yeonjun Choi¹*, Seyeon Jeong^{1*}, Jong Wook Kim²†, Beakcheol Jang^{1†}

¹Graduate School of Information, Yonsei University, Seoul, South Korea

²Department of Computer Science, Sangmyung University, Seoul, South Korea
{chlduswns99, seyeonjeong, bjang}@yonsei.ac.kr jkim@smu.ac.kr

Abstract

Large language models with retrieval-augmented generation systematically underutilize evidence in the middle of multi-document contexts, a phenomenon known as Lost in the Middle. Existing mitigation strategies, such as document reordering and attention steering, can reduce boundary bias but remain brittle in noisy settings, where inaccurate relevance estimation and spurious documents still obscure useful evidence in central positions. We propose DuQUAD, a multi-document QA framework that mitigates positional bias through dual-view reasoning and quality-aware evidence pruning. DuQUAD combines a Local Agent guided by a Structural Fusion Score with a Global Agent that reasons over the full document set, providing both targeted evidence recovery and broad contextual understanding. The framework further filters candidate evidence using explicit quality scoring and refines it at the sentence level to reduce noise. DuQUAD consistently outperforms strong Lost-in-the-Middle mitigation baselines, including hybrid retrieval and reranking, multi-agent reasoning, and context optimization methods, improving accuracy by 3.0% on average over the strongest competing baseline and gold evidence recall by 9.4%, while substantially reducing computational cost compared to existing multi-agent frameworks.

1 Introduction

Large Language Models (LLMs) have demonstrated strong performance on knowledge-intensive tasks (Achiam et al., 2023; Yang et al., 2025; Grattafiori et al., 2024), but remain limited by their static parametric knowledge, leading to issues such as obsolescence and hallucination. To address this, many approaches provide LLMs with external documents as additional context. However, even when

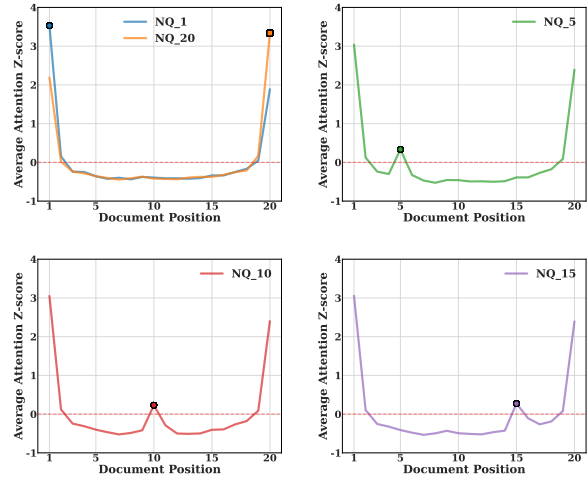


Figure 1: Attention z-score distribution across document positions on the Natural Questions (NQ) benchmark (NQ-1 to NQ-20, where NQ- n denotes instances in which the gold document appears at the n -th position).

relevant documents are available, LLMs often fail to use them effectively in multi-document settings due to a phenomenon known as Lost in the Middle (LiTM) (Liu et al., 2024). LLMs exhibit strong positional bias, disproportionately attending to documents at the boundaries while underutilizing evidence in central positions, which degrades multi-document QA performance (Liu et al., 2024; Wang et al., 2024; Hutter et al., 2025). This degradation is further amplified by contextual noise. We use *noise* to denote answer-irrelevant or weakly relevant documents in the retrieved set that disperse attention away from answer-bearing evidence and introduce distracting cues during reasoning.

Consistent with the U-shaped positional bias (Hsieh et al., 2024), attention concentrates at the beginning and end of the context. When gold documents appear in intermediate positions, localized spikes form a W-shaped pattern (Figure 1), but boundary attention still dominates, leaving mid-context evidence underutilized.

*Equal contribution.

†Co-corresponding authors.

To mitigate positional bias, most existing approaches rely on document reordering (Peysakhovich and Lerer, 2023; Liu et al., 2024; Jiang et al., 2024) or attention modification during training or inference (Hsieh et al., 2024; Jiang et al., 2024). However, relying on weakly calibrated signals such as semantic similarity or raw attention renders these approaches vulnerable to noise, particularly when gold evidence is buried among weakly relevant distractors at mid-context positions.

Motivated by these localized spikes, we propose a framework that mitigates positional bias through dual-view reasoning and quality-aware pruning. We operationalize the W-shaped attention pattern as an attention contrast signal and integrate 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, comparing against recent baselines including hybrid retrieval and reranking, LiTM mitigation strategies, and multi-agent frameworks. 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 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 identification and re-ranking of documents for local evidence selection.
- 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 filter noise and prevent hallucination.
- Extensive experiments across three multi-document QA benchmarks show that

DuQUAD improves accuracy by 3.0% on average over the strongest competing baseline, achieves a 13.8% gain over the stronger multi-agent baseline on NQ-15, and raises gold evidence recall by 9.4%.

2 Related Work

Positional Bias in Large Language Models

LLMs exhibit strong positional biases on multi-document inputs, tending to over-attend to tokens near the input boundaries while under-utilizing information in intermediate positions (Liu et al., 2024; Hutter et al., 2025). This phenomenon, known as Lost in the Middle (LiTM), is attributed to attention dynamics such as attention sinks and recency effects (Xiao et al., 2024; Gu et al., 2025). To mitigate these effects, prior work has proposed attention calibration and sorting techniques (Peysakhovich and Lerer, 2023; Hsieh et al., 2024), as well as positional scaling and dimension-level bias correction (Yu et al., 2024). However, these methods primarily target attention-level symptoms rather than how positional bias degrades evidence utilization in multi-document reasoning.

Training-based Approaches to LiTM Mitigation

Training-based methods address positional bias through fine-tuning or additional supervision (Liu et al., 2024; Han et al., 2024). Position-agnostic training and multi-step QA supervision encourage models to prioritize content relevance over absolute position (He et al., 2024; Pan et al., 2024), while full-context utilization training directly optimizes for mid-context evidence use (An et al., 2024). Although effective, these approaches require substantial model adaptation and cannot be applied to off-the-shelf models at inference time.

Training-free Approaches to LiTM Mitigation

Training-free methods mitigate positional bias at inference time without modifying model parameters (Wang et al., 2024). Document reordering and reranking place relevant passages at salient positions prior to reasoning (Peysakhovich and Lerer, 2023; Wang et al., 2025), but do not address suppression during reasoning itself. Context compression reduces input noise (Jiang et al., 2024; Verma, 2024), but risks discarding weak signals necessary for multi-hop reasoning. These methods do not explicitly model how positional bias suppresses evidence during reasoning, leaving mid-context evidence vulnerable to being overlooked.

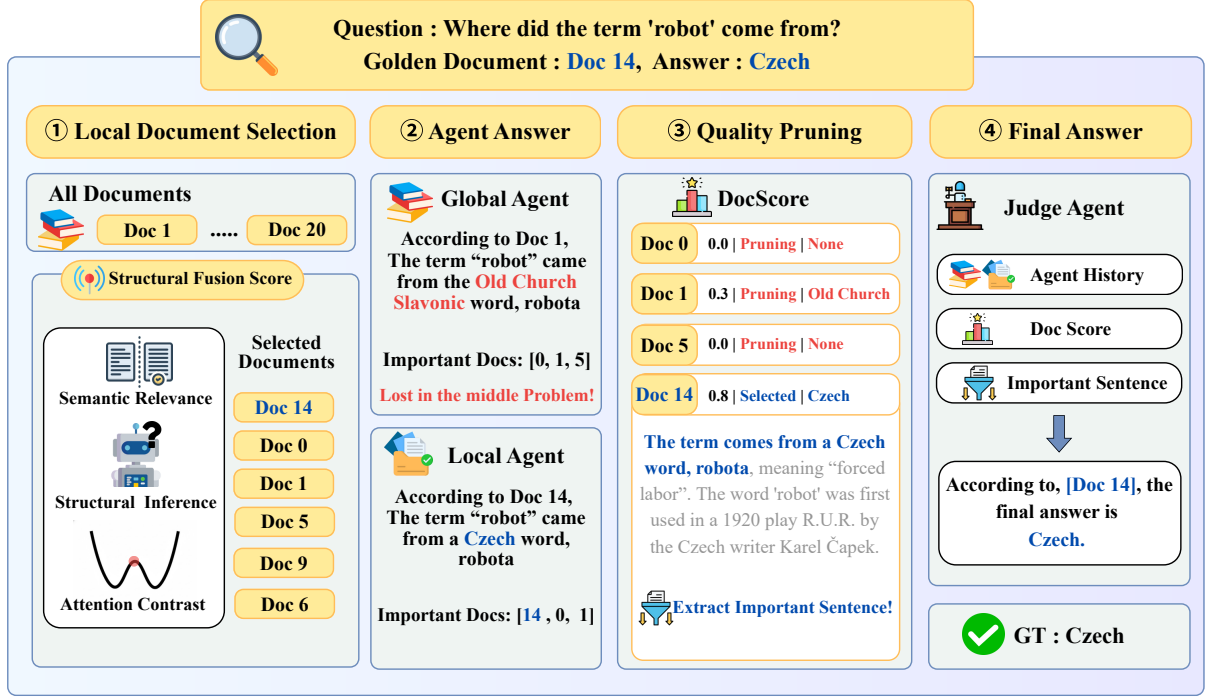


Figure 2: Overview of a dual-view multi-agent reasoning framework based on quality-aware evidence pruning. Given a question and a set of retrieved documents, the Structural Fusion Score (SFS) prioritizes a subset for the Local Agent, while the Global Agent operates on the full context. Selected documents 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.

Multi-agent Reasoning for Multi-document QA

Multi-agent reasoning extends single-agent LLMs by decomposing complex tasks into co-ordinated sub-problems, enabling role specialization and iterative refinement through agent collaboration. Prior work has explored various paradigms, including debate-based reasoning (Du et al., 2024) and role-based decomposition (e.g., planner–solver–verifier) (Chen et al., 2025). Building on these ideas, structured approaches such as Tree of Agents (Yu et al., 2025) and MAIN-RAG (Chang et al., 2025) organize reasoning and retrieval into hierarchical or multi-stage pipelines. However, these methods primarily focus on task decomposition and document selection, and do not explicitly account for positional bias, leaving mid-context evidence still under-utilized.

3 Methodology

We address reasoning failures in multi-document QA: despite accurate retrieval, standard 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 Figure 2.

3.1 Structural Fusion Score (SFS)

To identify gold documents under positional noise, we propose the Structural Fusion Score (SFS). While attention over multi-document inputs 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 capture query–document alignment, while attention contrast emphasizes localized spikes robust to positional bias.

3.1.1 Semantic Relevance (Φ_{sim})

To capture the semantic alignment between a query and documents, we use General Text Embeddings (GTE) (Li et al., 2023)¹, a strong open-source embedding model with stable semantic similarity performance. 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:

¹<https://huggingface.co/Alibaba-NLP/gte-large-en-v1.5>

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

3.1.2 Structural Inference (Φ_{ppl})

To capture structural inference signals, we evaluate the causal dependency of the query $Q = (q_1, \dots, q_{|Q|})$ on a document D_i by measuring the likelihood of generating it conditioned on the document context. In all experiments, Φ_{ppl} is computed using the same backbone LLM as the main reasoning model, which avoids introducing a separate scoring model and reduces cross-model inconsistency. This metric reflects intrinsic generative support beyond static embeddings.

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

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.

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 across multiple documents are heavily entangled with positional bias, sharp local deviations more reliably indicate gold evidence, revealing characteristic W-shaped patterns.

a_i is the average attention mass assigned from the final query token to the tokens of document D_i , aggregated over all layers and heads. Appendix J details this query-aware input construction and attention extraction procedure. Let μ_A and σ_A be the mean and standard deviation of a_i over the retrieved set $A = \{D_1, \dots, D_N\}$. 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. At document boundaries where symmetric neighbors are not available, we adopt a one-sided finite difference scheme, detailed in Appendix I.

3.1.4 Score 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:

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

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

Softmax addresses scale heterogeneity and induces 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 greater weight on structural inference signals. These weights are selected by grid search on Natural Questions based on average gold evidence recall across positions. As shown in Appendix E, the chosen configuration achieves the highest overall recall while remaining stable across nearby settings.

3.2 Dual-View Agent Collaboration

Although SFS prioritizes relevant evidence, using only a filtered subset may lose global context and cross-document dependencies. To address this, we introduce a Dual-View Agent Collaboration framework that reasons from two complementary views.

The Global Agent reasons over the complete retrieved document set, while the Local Agent focuses on a compact, noise-reduced subset re-ranked by SFS. This design provides mutual robustness: the Local Agent recovers evidence suppressed by positional bias or noise, and the Global Agent compensates for information lost through filtering. Both agents generate cited answers independently, which are aggregated into a unified reasoning log for position-resilient multi-document reasoning. We use $k = 6$ local documents based on a recall-accuracy trade-off analysis in Appendix D.

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 separates evidence selection from quality verification.

Important Document Selection Each agent independently identifies a set of important documents based on its own reasoning process, directly from its generated output. Specifically, each agent returns up to three document IDs as important documents, together with its reasoning results, following a predefined JSON output format (Appendix F, Figure 7).

Let $\mathcal{D}_G$ and $\mathcal{D}_L$ 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:

$$\mathcal{D}_{\text{imp}} = \mathcal{D}_G \cup \mathcal{D}_L \quad (4)$$

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 DocScore, an LLM-based document reliability score defined in this work, where $p_i \in [0, 1]$ estimates whether the question can be answered from the given document alone. DocScore is produced by a dedicated Score Agent with a fixed prompt, which returns a JSON-formatted confidence score together with an answer span (Appendix F, Figure 9).

We retain only high-confidence documents using a threshold-based rule:

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

We fix the threshold at $\tau = 0.4$ for all experiments. This value is selected based on the threshold sensitivity analysis in Appendix C (Table 5), where $\tau = 0.4$ is the highest threshold that preserves recall while still enabling meaningful pruning. Lower thresholds admit more irrelevant or weakly relevant documents, whereas higher thresholds risk discarding gold evidence too early.

For each retained document, we generate a document-specific provisional answer and select the top two sentences as supporting evidence based on word or keyword overlap between the provisional 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 supports final reasoning.

3.4 Final Decision via Judge Agent

The Judge Agent produces the final answer with a fixed prompt (Appendix F, Figure 8), 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 $\mathcal{H}_G$ and $\mathcal{H}_L$, 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:

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

By aggregating complementary agent outputs and evidence signals, the Judge Agent avoids over-reliance on any single source and supports robust answer generation. The complete DuQUAD pipeline is summarized in Algorithm 1.

4 Experiments

4.1 Experimental Setup

Backbone Model We use Llama-3.1-8B-Instruct (Grattafiori et al., 2024)² and Qwen2.5-7B-Instruct (Yang et al., 2024)³ as backbone models. Both provide sufficient context windows 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 by default.

Datasets We evaluate DuQUAD on three multi-document QA benchmarks. For Natural Questions (NQ) (Kwiatkowski et al., 2019), we follow the controlled positional bias protocol of (Liu et al., 2024), placing one gold document at position $n \in \{1, 5, 10, 15, 20\}$ among 19 distractors. MuSiQue (Trivedi et al., 2022) requires connected reasoning across 2–4 hops, where each question provides 2–4 supporting paragraphs among distractors and no single paragraph suffices to answer independently. HotpotQA (Yang et al., 2018) requires

²<https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct>

³<https://huggingface.co/Qwen/Qwen2.5-7B-Instruct>

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	<u>19.9</u>	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
	Jina-Reranker	69.0	68.4	69.2	69.4	68.6	29.3	17.6	27.0	<u>77.3</u>	51.6	67.4
	Qwen3-Emb.	69.3	69.3	69.3	69.3	69.3	34.7	18.0	27.0	76.8	50.3	67.2
	OpenAI	70.2	70.5	71.4	69.9	70.5	34.7	18.3	<u>27.7</u>	76.7	52.5	<u>68.9</u>
	Hybrid (RRF)	71.9	<u>72.0</u>	<u>72.1</u>	<u>72.2</u>	<u>72.3</u>	28.0	16.9	25.2	77.1	<u>53.8</u>	68.7
	+ Rerank	71.4	71.4	71.4	71.4	71.4	28.8	17.8	26.8	77.2	54.4	68.8
	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	<u>75.3</u>	67.0	65.6	64.7	65.4	<u>38.6</u>	12.0	22.4	76.9	37.0	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	<u>21.3</u>	<u>29.0</u>	79.2	<u>57.5</u>	<u>71.1</u>
	BM25	63.3	63.5	63.8	63.3	63.6	29.2	18.6	26.6	80.1	56.8	69.5
	Jina-Reranker	69.0	68.4	68.1	68.1	68.4	29.4	20.2	28.1	78.2	56.7	70.2
	Qwen3-Emb.	65.2	65.2	65.2	65.2	65.2	27.7	19.5	27.1	79.3	56.3	69.9
	OpenAI	65.4	65.4	65.9	65.0	65.2	30.2	20.0	27.5	<u>80.3</u>	57.2	<u>71.1</u>
	Hybrid (RRF)	67.5	67.5	67.4	67.5	67.7	28.4	19.5	27.3	79.0	55.1	68.4
	+ Rerank	69.9	<u>69.9</u>	<u>69.9</u>	<u>69.9</u>	<u>69.9</u>	29.9	19.9	27.7	77.8	56.9	70.2
	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	<u>73.6</u>	61.0	59.0	58.5	58.7	<u>32.4</u>	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	22.4	31.0	82.0	60.7	76.0

Table 1: Main experimental results comparing Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct. The highest score in each column is shown in **bold**, and the second-highest score is underlined.

bridge or comparison reasoning, evaluated in the distractor setting with two gold paragraphs among eight distractors. Detailed dataset characteristics, including input context statistics, are provided in Appendix B (Table 4), while per-experiment sample sizes are summarized in Table 3.

Evaluation Metrics Following Lost in the Middle (Liu et al., 2024), we evaluate performance using accuracy, defined as whether the generated response contains the ground-truth answer string. This metric is commonly used for multi-document QA with free-form outputs, focusing on answer presence rather than exact matching. 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 and reranking meth-

ods including SBERT (Reimers and Gurevych, 2019), BM25 (Crestani et al., 1998), OpenAI embeddings (OpenAI, 2024), Jina-Reranker (Wang et al., 2025), Qwen3-Embedding (Zhang et al., 2025), and two hybrid variants: Hybrid (RRF), which combines BM25 and Qwen3-Embedding rankings using Reciprocal Rank Fusion (RRF) (Cormack et al., 2009), and Hybrid (RRF) + Rerank, which further applies Jina-Reranker to the fused ranking; (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 such as Found in the Middle (FiM) (Hsieh et al., 2024) and LongLLMLingua (Jiang et al., 2024). Implementation details are provided in Appendix A.

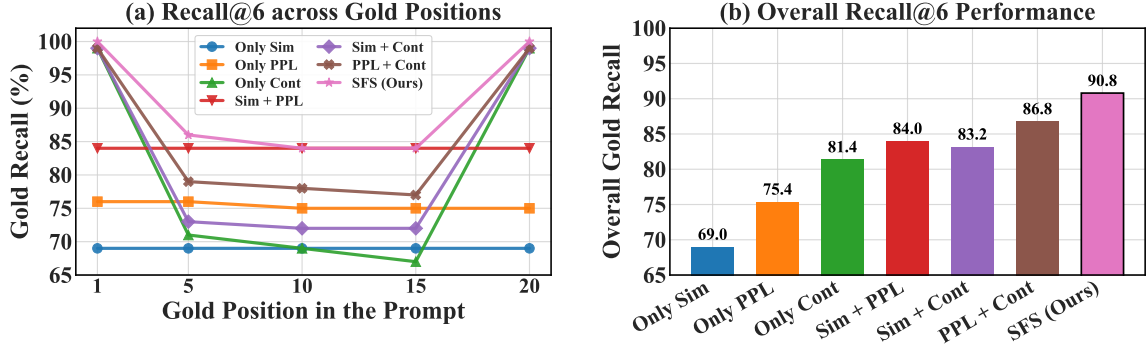


Figure 3: Impact of Score Fusion on Robustness. Sim, PPL, and Cont denote the three SFS signals: Semantic Relevance, Structural Inference, and Attention Contrast, respectively. (a) Recall of the gold document by gold position under the Top-6 setting. (b) Overall recall aggregated across all gold positions in (a). Computed on 100 NQ examples per gold position.

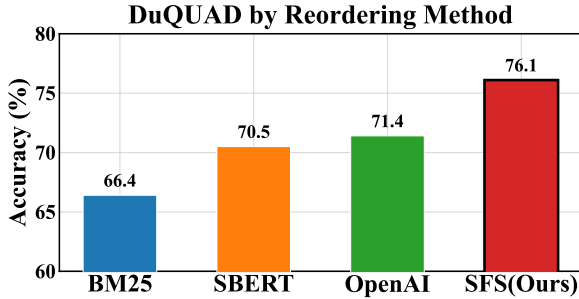


Figure 4: Impact of representative local document selection scores within DuQUAD on NQ-15. We compare sparse and dense retrieval-based alternatives against the proposed Structural Fusion Score (SFS) in terms of answer accuracy.

5 Results

5.1 Overall Performance

Table 1 reports accuracy on NQ, MuSiQue, and HotpotQA, with EM and F1 scores additionally reported for multi-hop tasks.

On NQ, the Vanilla LLM shows strong positional sensitivity, with accuracy varying substantially across gold-document positions. Reordering and reranking baselines reduce this positional variation, with the hybrid variants showing particularly stable performance across positions. Nevertheless, DuQUAD consistently achieves the highest accuracy across all NQ settings under both backbones. Context optimization and multi-agent baselines also remain challenged by LiTM: FiM degrades substantially at intermediate positions, while ToA and MAIN-RAG remain below DuQUAD despite their more complex reasoning pipelines. These results suggest that stronger document ranking, repositioning, or reasoning decomposition alone

does not fully address positional underutilization in multi-document reasoning.

Overall, DuQUAD achieves the best accuracy across all 14 accuracy settings in Table 1. It improves over the strongest competing baseline in each setting by 3.0 percentage points on average. The same trend holds for both Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct, suggesting that the gains are not specific to a single backbone.

On MuSiQue and HotpotQA, DuQUAD also achieves the best accuracy under both backbones. With Llama-3.1-8B-Instruct, it improves over the strongest competing baseline by 2.5 percentage points on MuSiQue and 2.9 percentage points on HotpotQA. DuQUAD further achieves the highest F1 on both multi-hop benchmarks. The only metric-level exception under the Llama backbone is HotpotQA EM, where Hybrid (RRF) + Rerank achieves 54.4%, compared with 52.6% for DuQUAD. This narrow EM advantage is consistent with HotpotQA’s relatively compact evidence setting, where strong document ordering can place the two gold paragraphs at salient positions, while strict exact-match scoring may penalize answers that are semantically correct but differ in surface form. With Qwen2.5-7B-Instruct, DuQUAD achieves the highest accuracy, EM, and F1 on both multi-hop benchmarks. These results demonstrate that DuQUAD maintains robust performance across both controlled positional settings and multi-hop reasoning tasks. Paired bootstrap resampling further shows that DuQUAD’s gains are statistically significant in 35 out of 36 comparisons ($p < 0.05$), with detailed results in Appendix G.1.

Method	NQ-15	MuSiQue	HotpotQA
w/o Local Agent	69.8	38.1	79.1
w/o Global Agent	71.1	36.1	80.0
w/o Quality Pruning	71.3	34.6	79.8
DuQUAD	76.1	41.1	80.2

Table 2: Effect of the Local Agent, Global Agent, and Quality-Aware Pruning. Ablation analysis across three datasets evaluating the impact on accuracy when omitting each component compared to the complete DuQUAD method.

5.2 Gold Evidence Recall and DocScore Stability

We evaluate local document selection using gold evidence recall, defined as whether the gold document is included in the candidate set selected for the Local Agent. As shown in Figure 3, Attention Contrast achieves the highest recall among single-signal approaches at 81.4%, surpassing Semantic Relevance with 69.0% and Structural Inference 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 Structural Inference and Attention Contrast, and a 9.4% gain over the best single-signal method. These results indicate that integrating complementary signals enables more robust identification of gold evidence across different document positions.

DocScore further provides a stable quality signal for evidence pruning. Across NQ-1/5/10/15/20, gold documents consistently receive higher DocScores than non-gold documents and are selected as Important Documents more frequently under both Llama and Qwen backbones. Cross-model analysis further shows 90.4% agreement on gold evidence selection, suggesting that DocScore captures a stable answerability signal rather than backbone-specific behavior. Detailed statistics are provided in Appendix L.

5.3 Ablation Analysis

Table 2 reports ablation results evaluating the contributions of the Local Agent, the Global Agent, and quality-aware evidence pruning. Removing any component consistently degrades performance, confirming that DuQUAD’s gains come from the interaction of complementary reasoning views and explicit evidence filtering. The largest drop on NQ-15 occurs when the Local Agent is removed, indicating that global context reason-

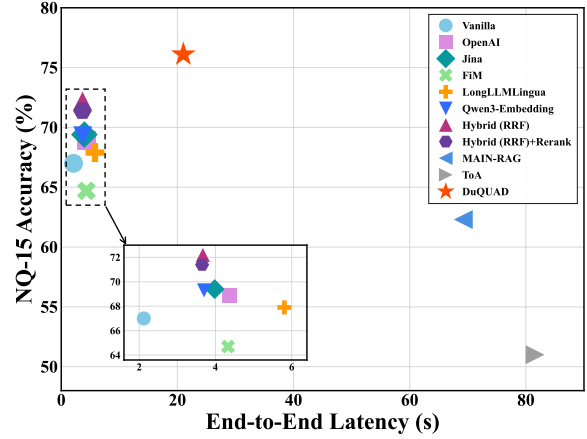


Figure 5: Accuracy–latency trade-off of representative baselines on NQ-15. The inset magnifies the low-latency region for clearer comparison among retrieval, reranking, and context optimization methods.

ing alone is insufficient when answer-bearing evidence is embedded among many distractors. In contrast, MuSiQue shows larger drops when the Global Agent or quality-aware pruning is removed, suggesting that multi-hop reasoning requires both cross-document context preservation and filtering of noisy or weakly aligned reasoning traces. The smaller changes on HotpotQA suggest less headroom under its smaller candidate set and more explicit supporting facts.

Figure 4 further analyzes document prioritization under the NQ-15 setting using representative local document selection scores. Among the evaluated alternatives, OpenAI achieves the highest accuracy at 71.4%, compared with 76.1% for the proposed Structural Fusion Score (SFS). These results show that replacing SFS with standard sparse or dense retrieval scores reduces performance, supporting the benefit of combining semantic relevance, structural inference, and attention contrast for local document prioritization. More broadly, DuQUAD integrates this SFS-based local prioritization with dual-view reasoning and quality-aware pruning. Appendix K provides quantitative disagreement analysis and qualitative examples of Local–Global complementarity and joint failure cases.

5.4 Cost Efficiency Analysis

We analyze the cost-efficiency of DuQUAD with respect to end-to-end latency on NQ-15. Figure 5 shows the accuracy–latency trade-off across representative baselines. Among lightweight retrieval and reranking methods, Hybrid (RRF) provides

the strongest accuracy–latency trade-off, achieving 72.2% accuracy with an end-to-end latency of 3.67 s. DuQUAD further improves accuracy to 76.1%, at the cost of additional inference latency.

Compared with existing multi-agent frameworks, however, DuQUAD is both more accurate and substantially more efficient: it improves accuracy by 13.8% over MAIN-RAG and 25.1% over ToA while reducing end-to-end latency by $3.3\times$ and $3.9\times$, respectively. These results show that DuQUAD incurs additional overhead relative to lightweight retrieval and reranking methods, but achieves higher accuracy while remaining substantially more efficient than existing multi-agent frameworks. As the Local and Global Agents operate independently, they can also be executed in parallel to further reduce latency. Detailed cost statistics are provided in Appendix H.

6 Conclusion

We investigated the limitations of multi-document QA under positional bias, specifically the Lost in the Middle phenomenon. To address it, we propose DuQUAD, a dual-view reasoning framework that pairs a Local Agent operating on a prioritized evidence subset with a Global Agent that reasons over the full document set. Local selection is guided by the Structural Fusion Score, which integrates semantic relevance, structural inference, and attention contrast to robustly identify gold documents despite positional noise. Quality-aware pruning further removes low-confidence evidence before final reasoning, and together these components raise gold evidence recall to 90.8%. As a result, DuQUAD mitigates positional bias and achieves the best accuracy among the evaluated baselines across all benchmarks, while requiring substantially less computation than existing multi-agent frameworks. These results demonstrate that DuQUAD offers a favorable balance between accuracy and cost for robust reasoning in multi-document QA under positional bias.

7 Limitations

While DuQUAD effectively mitigates positional bias in multi-document QA settings, it also has several limitations. First, DuQUAD employs an explicit evidence pruning mechanism to suppress noisy or low-confidence documents. Although this improves robustness against positional noise, it may reduce contextual coverage when relevant ev-

idence is weakly expressed or distributed across multiple documents, particularly for complex multi-hop questions.

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.

Third, even when answer-bearing evidence is correctly retrieved and scored, the final synthesis stage may still fail in complex multi-hop questions where the reasoning chain must track closely related intermediate entities, as reflected in the quantitative and qualitative analyses in Appendix K.

Finally, our evaluation is limited to Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct; broader evaluation across model families, capabilities, and sizes would further assess the generalizability of DuQUAD.

These limitations reflect inherent trade-offs in multi-document QA systems that balance contextual coverage, noise suppression, and reasoning stability.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) under Grant No. RS-2023-00273751 and by the Institute of Information Communications Technology Planning Evaluation (IITP) under Grant No. RS-2024-00397085, funded by the Korean government.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, Jian-Guang Lou, and Weizhu Chen. 2024. Make your llm fully utilize the context. *Advances in Neural Information Processing Systems*, 37:62160–62188.
- 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.

- Justin Chen, Archiki Prasad, Swarnadeep Saha, Elias Stengel-Eskin, and Mohit Bansal. 2025. [MAgICoRe: Multi-agent, iterative, coarse-to-fine refinement for reasoning](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32663–32686, Suzhou, China. Association for Computational Linguistics.
- Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. 2009. [Reciprocal rank fusion outperforms condorcet and individual rank learning methods](#). In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 758–759. ACM.
- 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.
- Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multi-agent debate. In *Forty-first international conference on machine learning*.
- Andrew Estornell and Yang Liu. 2024. Multi-llm debate: Framework, principals, and interventions. *Advances in Neural Information Processing Systems*, 37:28938–28964.
- Rafael C Gonzalez. 2009. *Digital image processing*. Pearson education india.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). Preprint, arXiv:2407.21783.
- Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. 2025. [When attention sink emerges in language models: An empirical view](#). In *International Conference on Learning Representations*, volume 2025, pages 97114–97144.
- Chi Han, Qifan Wang, Hao Peng, Wenhan Xiong, Yu Chen, Heng Ji, and Sinong Wang. 2024. Lm-infinite: Zero-shot extreme length generalization for large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3991–4008.
- Junqing He, Kunhao Pan, Xiaoqun Dong, Zhuoyang Song, Yibo Liu, Qianguo Sun, Yuxin Liang, Hao Wang, Enming Zhang, and Jiaying Zhang. 2024. [Never lost in the middle: Mastering long-context question answering with position-agnostic compositional training](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13628–13642, Bangkok, Thailand. Association for Computational Linguistics.
- 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.
- Jan Hutter, David Rau, Maarten Marx, and Jaap Kamps. 2025. Lost but not only in the middle: Positional bias in retrieval augmented generation. In *European Conference on Information Retrieval*, pages 247–261. Springer.
- 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.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. [Natural questions: A benchmark for question answering research](#). *Transactions of the Association for Computational Linguistics*, 7:452–466.
- 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.
- OpenAI. 2024. Text embedding 3 models. <https://platform.openai.com/docs/models/text-embedding-3-large>. Accessed: 2025-12-25.
- Ruotong Pan, Boxi Cao, Hongyu Lin, Xianpei Han, Jia Zheng, Sirui Wang, Xunliang Cai, and Le Sun. 2024. Not all contexts are equal: Teaching llms credibility-aware generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19844–19863.

- Alexander Peysakhovich and Adam Lerer. 2023. Attention sorting combats recency bias in long context language models. *arXiv preprint arXiv:2310.01427*.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- 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.
- Sourav Verma. 2024. Contextual compression in retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2409.13385*.
- Feng Wang, Yuqing Li, and Han Xiao. 2025. jina-reranker-v3: Last but not late interaction for document reranking. *arXiv preprint arXiv:2509.25085*.
- Minzheng Wang, Longze Chen, Fu Cheng, Shengyi Liao, Xinghua Zhang, Bingli Wu, Haiyang Yu, Nan Xu, Lei Zhang, Run Luo, and 1 others. 2024. Leave no document behind: Benchmarking long-context llms with extended multi-doc qa. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5627–5646.
- Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. [Efficient streaming language models with attention sinks](#). In *International Conference on Learning Representations*, volume 2024, pages 21875–21895.
- 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*.
- 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, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.
- 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](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 4574–4592, Suzhou, China. Association for Computational Linguistics.
- Yijiong Yu, Huiqiang Jiang, Xufang Luo, Qianhui Wu, Chin-Yew Lin, Dongsheng Li, Yuqing Yang, Yongfeng Huang, and Lili Qiu. 2024. Mitigate position bias in large language models via scaling a single dimension. *arXiv preprint arXiv:2406.02536*.
- Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, and 1 others. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

A Baseline Implementation Details

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

(2) **Reordering and Reranking.** (i) SBERT Reordering (Reimers and Gurevych, 2019) is 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) is 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 using proprietary embeddings. (iv) Jina-Reranker (Wang et al., 2025) is a reranking baseline based on cross-encoder-style query–document interaction. Candidate documents are re-ranked according to their relevance scores, enabling more precise ordering than bi-encoder similarity alone. (v) Qwen3-Embedding (Zhang et al., 2025) is a dense embedding baseline using a recent open-source embedding model. Documents are rearranged based on cosine similarity between query and document embeddings. (vi) Hybrid (RRF) combines the BM25 and Qwen3-Embedding rankings using Reciprocal Rank Fusion (RRF) (Cormack et al., 2009). RRF assigns each document a fused score by summing reciprocal rank-based scores across the two ranked lists:

$$\text{RRF}(d) = \sum_{m \in \{\text{BM25, Dense}\}} \frac{1}{k + r_m(d)}, \quad (7)$$

where $r_m(d)$ denotes the rank of document d under retrieval method m . We use $k = 60$ and reorder

Experiment	Dataset	Samples	Reference
W-shape Visualization	NQ	100×5	Figure 1
Main Experiment	NQ	$1,000 \times 5$	Table 1
	MuSiQue	1,000	
	HotpotQA	1,000	
30-Document Robustness	NQ-15	1,000	Table 8
Gold Recall Analysis	NQ	100×5	Figure 3
DocScore / Important-Doc Statistics	NQ	$1,000 \times 5$	Table 11
Cross-Model Evidence Stability	NQ	$1,000 \times 5$	Table 12
Component Ablation	NQ-15	1,000	Table 2
	MuSiQue	1,000	
	HotpotQA	1,000	
	NQ	1,000	
Local-Global Disagreement	MuSiQue	1,000	Table 10
	HotpotQA	1,000	
Local-Selection Ablation	NQ-15	1,000	Figure 4
Top- k Analysis	NQ-15	1,000	Figure 6
Cost Analysis	NQ-15	1,000	Table 9

Table 3: Summary of experimental settings and datasets. For NQ experiments evaluated over multiple gold positions, the sample count is shown as (examples per position) $\times$ (number of positions); the five positions are $n \in \{1, 5, 10, 15, 20\}$.

Dataset	Samples	Docs	Gold Docs	Question Type	Input Tokens		
					Mean	Std.	Max
NQ	1,000	20	1	Single-hop factoid	2,781	140	3,396
MuSiQue	1,000	20	2-4	2-4 hop	2,446	495	4,751
HotpotQA	1,000	10	2	2-hop	1,351	349	3,220
NQ (30 docs)	1,000	30	1	Single-hop factoid	4,239	205	5,474

Table 4: Dataset characteristics and input context statistics. Input lengths include all documents and the question and are measured using the Llama-3.1-8B-Instruct tokenizer. For NQ, each positional split contains 1,000 examples; the 30-document setting is an additional robustness evaluation with the gold document at position 15.

documents according to the resulting fused scores, with Qwen3-Embedding serving as the dense retriever. (vii) Hybrid (RRF) + Rerank further applies Jina-Reranker to the RRF-fused ranking, combining sparse and dense retrieval signals with cross-encoder reranking.

(3) 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) is a collaborative architecture with specialized agents. It employs adaptive filtering to dynamically set inclusion thresholds, retaining high-confidence evidence while discarding noise.

(4) Context Optimization. (i) Found in the Middle (FiM) (Hsieh et al., 2024) is a position-aware strategy for mitigating the Lost in the Middle phenomenon. It applies a U-shaped reordering strat-

egy that places the most relevant documents at the beginning and end of the context. (ii) LongLLM-Lingua (Jiang et al., 2024) focuses on prompt compression. Following the original implementation settings, it utilizes Llama-2-7b-chat-hf⁴ to estimate token perplexity, discarding low-importance tokens to compress the prompt while preserving key information.

B Experimental Settings and Dataset Details

This appendix describes the datasets and experimental configurations used throughout the paper. Table 4 summarizes the main characteristics of each benchmark, including the number of evaluation samples, documents, gold documents, question type, and input context length statistics.

⁴<https://huggingface.co/meta-llama/Llama-2-7b-chat-hf>

Natural Questions (NQ). The original NQ corpus (Kwiatkowski et al., 2019) contains 307,373 training examples derived from real Google search queries, where annotators mark long answers and short answers on Wikipedia pages. In our evaluation, we use the controlled positional bias protocol of (Liu et al., 2024) rather than the original reading comprehension task. Each instance places one gold document at a designated position among 19 distractors within a 20-document sequence. Five positional splits are constructed for $n \in \{1, 5, 10, 15, 20\}$, with 1,000 examples per split, and all questions are single-hop. The five splits share the same document sets and differ only in the position of the gold document.

MuSiQue. MuSiQue (Trivedi et al., 2022) is constructed via bottom-up composition of single-hop questions drawn from five seed datasets into connected multi-hop chains. The dataset contains approximately 25K questions spanning 2–4 hops across six reasoning graph types. Each question enforces that one reasoning step critically depends on another, and no single paragraph suffices for an independent answer, preventing the disconnected reasoning shortcuts observed in prior benchmarks. We use instances from the answerable subset, each with its original supporting and distractor paragraphs.

HotpotQA. HotpotQA (Yang et al., 2018) contains 112,779 Wikipedia-based question–answer pairs annotated with sentence-level supporting facts. Each question requires 2-hop reasoning and is categorized as either bridge, where an intermediate entity from one paragraph is needed to locate the answer in another, or comparison, which requires contrasting properties across two paragraphs. We evaluate in the distractor setting, where two gold paragraphs are presented alongside eight distractors.

Table 4 summarizes the benchmark-level evaluation characteristics and context length statistics, while the datasets and sample sizes used for individual experiments vary depending on the experimental purpose. All experiment-specific settings are summarized in Table 3.

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 ex-

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.

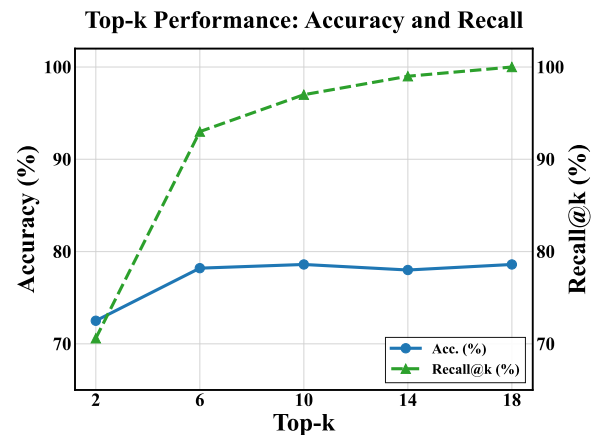


Figure 6: Impact of the number of documents provided to the Local Agent on Accuracy and Recall@k.

pected 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 Recall–Accuracy Trade-off for Local Document Selection

Figure 6 shows the effect of the local document budget k on NQ-15. Recall@ k increases monotonically as more documents are included and begins to saturate around $k = 6$. Answer accuracy exhibits a similar trend, improving substantially up to $k = 6$ while showing only marginal gains or

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 gold 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.

slight fluctuations beyond this point. This suggests that additional documents beyond $k = 6$ mainly increase contextual noise rather than improving final answer quality. We therefore use $k = 6$ as the default local-document budget, balancing high recall with controlled noise.

E 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 gold 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.

F 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 7, 8, and 9, respectively, without modification. In particular, the reasoning agent prompt in Figure 7 in-

structs each agent to answer the question using only the provided documents and to return the top 3 important document IDs in a fixed JSON format, which is later used to construct the candidate document pool. The score estimation prompt in Figure 9 defines the DocScore procedure: for each selected candidate document, the Score Agent estimates whether the question can be answered from that document alone and returns a JSON object containing a confidence score $p \in [0, 1]$ together with the best answer span or phrase.

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 paper⁵. 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.

Artifact use. All existing artifacts used in this work are used solely for research and evaluation purposes, consistent with their intended use.

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.

G Additional Evaluation Analyses

G.1 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

⁵<https://github.com/nelson-liu/lost-in-the-middle>

Dataset (DuQUAD Acc.)	Baseline	Acc. (%)	Diff. (95% CI)	p-value
NQ-15 (76.1)	Vanilla LLM	67.0	+9.1 (6.5 – 11.9)	< 0.0001***
	SBERT	70.1	+6.0 (3.3 – 8.8)	< 0.0001***
	BM25	69.2	+6.9 (4.1 – 9.7)	< 0.0001***
	Jina-Reranker	69.4	+6.7 (4.1 – 9.4)	< 0.0001***
	Qwen3-Embedding	69.3	+6.8 (4.0 – 9.6)	< 0.0001***
	OpenAI	69.9	+6.2 (3.4 – 9.0)	< 0.0001***
	Hybrid (RRF)	72.2	+3.9 (1.2 – 6.6)	0.0042**
	+ Rerank	71.4	+4.7 (2.1 – 7.3)	0.0004***
	ToA	51.0	+25.1 (22.0 – 28.4)	< 0.0001***
	MAIN-RAG	62.3	+13.8 (10.9 – 16.8)	< 0.0001***
	LongLLMLingua	67.9	+8.2 (5.5 – 11.1)	< 0.0001***
	FiM	64.7	+11.4 (8.6 – 14.2)	< 0.0001***
MuSiQue (41.1)	Vanilla LLM	33.1	+8.0 (5.4 – 10.6)	< 0.0001***
	SBERT	24.9	+16.2 (13.4 – 18.9)	< 0.0001***
	BM25	33.6	+7.5 (5.1 – 10.0)	< 0.0001***
	Jina-Reranker	29.3	+11.8 (9.2 – 14.5)	< 0.0001***
	Qwen3-Embedding	34.7	+6.4 (3.9 – 8.9)	< 0.0001***
	OpenAI	34.7	+6.4 (3.8 – 9.1)	< 0.0001***
	Hybrid (RRF)	28.0	+13.1 (10.5 – 15.8)	< 0.0001***
	+ Rerank	28.8	+12.3 (9.7 – 15.0)	< 0.0001***
	ToA	21.5	+19.6 (16.6 – 22.6)	< 0.0001***
	MAIN-RAG	19.0	+22.1 (19.2 – 25.0)	< 0.0001***
	LongLLMLingua	30.8	+10.3 (7.5 – 13.2)	< 0.0001***
	FiM	38.6	+2.5 (-0.4 – 5.4)	0.0910
HotpotQA (80.2)	Vanilla LLM	73.8	+6.4 (3.9 – 9.0)	< 0.0001***
	SBERT	76.8	+3.4 (1.0 – 5.8)	0.0066**
	BM25	75.3	+4.9 (2.4 – 7.4)	< 0.0001***
	Jina-Reranker	77.3	+2.9 (0.4 – 5.3)	0.0242*
	Qwen3-Embedding	76.8	+3.4 (1.0 – 5.8)	0.0062**
	OpenAI	76.7	+3.5 (1.0 – 6.0)	0.0072**
	Hybrid (RRF)	77.1	+3.1 (0.5 – 5.7)	0.0228*
	+ Rerank	77.2	+3.0 (0.5 – 5.5)	0.0216*
	ToA	57.3	+22.9 (19.8 – 26.1)	< 0.0001***
	MAIN-RAG	60.5	+19.7 (16.7 – 22.7)	< 0.0001***
	LongLLMLingua	70.2	+10.0 (7.2 – 12.8)	< 0.0001***
	FiM	76.9	+3.3 (0.8 – 5.9)	0.0106*

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: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

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 a two-sided bootstrap hypothesis test under the null hypothesis that the expected difference is zero. Two-sided p-values are computed as twice the smaller empirical tail probability of the bootstrap distribution relative to zero. 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. DuQUAD yields positive mean improvements over all baselines across the 36 pairwise comparisons (12 baselines $\times$ 3 datasets), with 35 of them reaching statistical significance at $p < 0.05$. Significant improvements are observed across reordering and reranking methods, multi-agent frameworks, and context optimization techniques, in-

cluding strong baselines such as Hybrid (RRF), Hybrid (RRF) + Rerank, Jina-Reranker, Qwen3-Embedding, LongLLMLingua, and FiM on NQ-15 and HotpotQA.

The single exception is DuQUAD vs. FiM on MuSiQue. Two factors underlie this result: FiM is a strong baseline on multi-hop reasoning, as its U-shaped reordering places relevant documents at salient positions well suited to compositional evidence, and the absolute gap (+2.5%) is the smallest among all MuSiQue comparisons, where the remaining gaps (+6.4% to +22.1%) yield clearer statistical separation. The direction of improvement nonetheless remains consistent, and DuQUAD shows clearer statistical advantages in settings with stronger positional bias, such as NQ.

G.2 Robustness to Increased Document Count

We further evaluate robustness to an increased number of documents by extending NQ-15 from 20 to 30 documents while keeping the gold document at position 15. Table 8 compares accuracy with the

Method	20 Docs	30 Docs	Diff.
Vanilla LLM	67.0	60.8	-6.2
OpenAI	69.9	66.1	-3.8
Hybrid (RRF)	72.2	64.8	-7.4
Hybrid (RRF) + Rerank	71.4	67.8	-3.6
FiM	64.7	62.7	-2.0
DuQUAD	76.1	75.6	-0.5

Table 8: Accuracy on NQ-15 under the 20- and 30-document settings. Diff. denotes the accuracy change when increasing the number of documents from 20 to 30.

corresponding 20-document NQ-15 setting.

As shown in Table 8, increasing the number of documents substantially degrades Vanilla LLM and several competing baselines, whereas DuQUAD remains stable, decreasing by only 0.5 points. Among the evaluated baselines in the 30-document setting, Hybrid (RRF) + Rerank achieves 67.8% accuracy, leaving a 7.8-point gap to DuQUAD’s 75.6%. FiM also exhibits a relatively small degradation of 2.0 points, but its absolute accuracy remains substantially lower at 62.7%.

H Cost Analysis

Table 9 reports the inference cost of DuQUAD and representative baselines in terms of LLM calls, total token consumption, and end-to-end latency, all measured on the same 1,000-instance NQ-15 evaluation set used for the NQ-15 accuracy results in Table 1 and normalized to the Vanilla setting. LLM calls reflect the number of model invocations in each pipeline, token counts capture the effective computation incurred when processing multi-document contexts, and end-to-end latency measures the wall-clock time required to produce a final answer. For retrieval and reranking methods, including OpenAI, Qwen3-Embedding, Jina-Reranker, and the hybrid variants, retrieval-side computations are performed by separate models rather than the answer-generation LLM, and their associated calls and tokens are therefore not included in the reported LLM call and token counts, while their computational overhead is reflected in end-to-end latency.

Table 9 shows that lightweight retrieval and reranking methods require only a single answer-generation LLM call, whereas existing multi-agent frameworks require substantially more calls and latency. Among the lightweight methods, Hybrid (RRF) achieves the highest NQ-15 accuracy at

72.2% with an end-to-end latency of 3.67 s, while Qwen3-Embedding and Hybrid (RRF) + Rerank show comparable latency at 3.70 s and 3.65 s, respectively. DuQUAD further improves accuracy to 76.1%, while using 4.10 LLM calls and incurring $9.93\times$ end-to-end latency relative to Vanilla.

Compared to MAIN-RAG, DuQUAD reduces LLM calls by $10\times$ (4.10 vs. 41.00), token consumption by $2.35\times$ (5,914 vs. 13,880), and end-to-end latency by $3.29\times$ (21.03s vs. 69.22s). Compared to ToA, DuQUAD reduces LLM calls by $5.65\times$, token consumption by $2.81\times$, and end-to-end latency by $3.88\times$. At the same time, DuQUAD outperforms both methods in accuracy across all three benchmarks (Table 1), indicating that the accuracy benefits of multi-agent reasoning can be retained while substantially reducing its characteristic cost overhead.

Although DuQUAD is more costly than lightweight retrieval and reranking methods due to dual-view reasoning and quality-aware pruning, its overhead remains moderate relative to existing multi-agent frameworks and is justified when accuracy is prioritized under positional bias and noise.

I 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, attention 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.

J Document-Level Attention Scoring

This section describes how we compute document-level attention scores, which are used as one component of the Structural Fusion Score.

Method	LLM Calls		Tokens		E2E Latency (s)	
	Value	Ratio	Value	Ratio	Value	Ratio
Vanilla LLM	1.00	1.00×	2863.50	1.00×	2.12	1.00×
OpenAI	1.00	1.00×	2900.31	1.01×	4.38	2.07×
Qwen3-Embedding	1.00	1.00×	2940.39	1.03×	3.70	1.75×
Jina-Reranker	1.00	1.00×	2940.53	1.03×	3.98	1.88×
Hybrid (RRF)	1.00	1.00×	2939.80	1.03×	3.67	1.73×
+ Rerank	1.00	1.00×	2934.68	1.02×	3.65	1.72×
FiM	2.00	2.00×	5437.09	1.90×	4.33	2.04×
LongLLMLingua	2.00	2.00×	5822.48	2.03×	5.81	2.74×
MAIN-RAG	41.00	41.00×	13880.14	4.85×	69.22	32.67×
ToA	23.17	23.17×	16613.13	5.80×	81.58	38.50×
DuQUAD	4.10	4.10×	5914.07	2.07×	21.03	9.93×

Table 9: Cost comparison normalized to Vanilla LLM in terms of LLM calls, tokens, and end-to-end latency.

Dataset	Both Correct	Global Only	Local Only	Both Wrong	Disagreement Rate	Judge Acc. on Disagreement
NQ	579	131	112	178	24.3%	74.5% (181/243)
HotpotQA	746	72	63	119	13.5%	58.5% (79/135)
MuSiQue	291	103	83	523	18.6%	67.2% (125/186)

Table 10: Local–Global agent agreement and disagreement statistics. Each dataset contains 1,000 evaluation examples. Disagreement denotes cases in which exactly one agent is correct, and Judge accuracy is measured only over these disagreement cases.

Input sequence construction. Given a question Q and retrieved documents $\{D_1, \dots, D_N\}$, we follow the query-aware contextualization format of Liu et al. (2024), which places the query before and after the document sequence:

$$[Q; D_1; D_2; \dots; D_N; Q].$$

The second occurrence of Q serves as the final query segment. In decoder-only LLMs, the last token of this final query segment can causally attend to all preceding document tokens. Each document D_i is preceded by a delimiter token (e.g., [DOC]), and its token span is recorded as $\text{span}(D_i) = [s_i, e_i]$.

Document-level attention score. Let t^* denote the position of the last token in the final query segment. We run the backbone LLM with attention outputs enabled and extract the self-attention vector from t^* to all previous input tokens. For each layer ℓ and head h , this vector is denoted by $A_{t^*,j}^{(\ell,h)}$. We average it over all layers and heads:

$$r[j] = \frac{1}{LH} \sum_{\ell=1}^L \sum_{h=1}^H A_{t^*,j}^{(\ell,h)}.$$

Here, $r[j]$ denotes the average attention mass assigned from the final query token to token position j . The document-level attention score for D_i is then computed as the mean attention mass over its token span:

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

This computation requires access to the model’s self-attention tensors, such as through `output_attentions=True` in open-source implementations. It does not modify model parameters or decoding behavior, but assumes that attention weights are available.

K Analysis of Local–Global Complementarity

To quantify the complementary behavior of the Local and Global Agents, we categorize each prediction according to whether each agent produces the correct answer. Table 10 reports the resulting statistics over 1,000 examples from each dataset. We define a disagreement case as one in which exactly one of the two agents is correct, and report Judge accuracy as the proportion of these cases for

Model	Document Type	DS Mean	ID Avg.
Llama	Gold Doc	0.647	85.7
	Non-Gold Doc	0.324	14.3
Qwen	Gold Doc	0.722	83.0
	Non-Gold Doc	0.348	17.0

Table 11: DocScore (DS) and Important Document (ID) statistics on Natural Questions (NQ-1/5/10/15/20). 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.

which the final Judge Agent produces the correct answer.

Table 10 reveals that both Global-only and Local-only successes occur consistently across all three datasets, indicating that neither view subsumes the other. The disagreement rate ranges from 13.5% on HotpotQA to 24.3% on NQ, and the Judge Agent resolves these disagreements correctly more often than not on all three datasets, reaching 74.5% accuracy on NQ, 58.5% on HotpotQA, and 67.2% on MuSiQue. These results quantitatively support the complementary roles of the two reasoning views and the benefit of integrating them at the final decision stage.

The remaining errors also reveal a distinct limitation on more complex multi-hop reasoning. On MuSiQue, both agents are incorrect on 523 of 1,000 examples (52.3%), substantially more often than on NQ or HotpotQA. This suggests that many MuSiQue failures arise before final adjudication, when both reasoning views fail to correctly track or integrate intermediate entities across multiple hops.

To complement this aggregate analysis, we present three representative cases that illustrate how the Local and Global Agents interact: two cases in which one agent fails while the other recovers the correct answer, and one failure case in which both agents converge on the same incorrect reasoning.

Case 1: Global Recovery under Local Error.

For the question “*who won the mens figure skating in 2018 olympics*”, the gold document is DOC 14, and the correct answer is Yuzuru Hanyu. The Global Agent correctly identifies the answer from the full document set by citing DOC 14. In contrast, the Local Agent predicts Nathan Chen, even though DOC 14 is included in its selected local subset. Instead of relying on the answer-bearing statement,

the Local branch is distracted by another document describing Nathan Chen’s technical achievements, which are salient but not sufficient to answer the question. During quality-aware pruning, DOC 14 receives a DocScore of 1.0, while the other candidate documents receive 0.0. The Judge Agent therefore rejects the Local Agent’s incorrect claim and produces the correct final answer, Yuzuru Hanyu. This case shows that even when the Local branch fails, the Global branch and subsequent pruning can prevent the error from propagating to the final decision.

Case 2: Local Recovery under Global Uncertainty.

We also observe cases in which the Global branch follows a more indirect reasoning path, whereas the Local branch recovers the answer more directly. For the question “*who performed the half-time show at super bowl 51*”, the Global Agent initially states that the performer is unclear and refers to surrounding information that Adele had been discussed as a possible candidate. Although it eventually reaches the correct answer by citing DOC 14, its reasoning does not immediately converge on the core answer-bearing evidence. In contrast, the Local Agent explicitly states that “*Lady Gaga performed the halftime show at Super Bowl LI (which is Super Bowl 51) [DOC 14]*,” directly identifying the correct entity and clarifying the equivalence between Super Bowl LI and Super Bowl 51. At the final integration stage, this complementary behavior allows the system to converge stably to the correct answer, Lady Gaga. This case shows that even when one branch becomes temporarily unstable, the other branch can preserve the correct answer signal and support a stable final prediction.

Case 3: Joint Failure under Multi-hop Entity Tracking.

We also observe failure cases in which DuQUAD retrieves several relevant documents but follows the wrong entity branch during final synthesis. In a MuSiQue example, the question requires a four-hop chain from Lion Air to Juanda International Airport, then to Indonesia, then to East Timor through the Indonesia–Timor Leste Commission of Truth and Friendship, and finally to the president of East Timor (Francisco Guterres). The Local Agent selects multiple gold documents, including DOC 9, which explicitly states “*President Francisco Guterres*,” and quality-aware pruning assigns high DocScores to the East Timor and commission documents (DOC 9: $p = 0.80$; DOC 14: $p = 0.90$). However, both the Local and Global Agents focus

Metric	Value
Gold evidence stability	90.4%
Distractor variability	90.1%

Table 12: Cross-model evidence selection stability on Natural Questions (NQ-1/5/10/15/20).

on the Indonesia branch of the reasoning chain and infer the Indonesian president, Susilo Bambang Yudhoyono. The Judge Agent then follows this shared but incorrect intermediate entity instead of resolving the target country as East Timor, and outputs the wrong final answer. This case shows that although DuQUAD reliably surfaces and scores answer-bearing evidence, complex multi-hop questions can still fail when the final synthesis stage must track closely related entities across several intermediate hops.

L DocScore Statistics and Cross-Model Evidence Stability

Table 11 summarizes the DocScore and Important Document statistics used in the main analysis. For both backbones, gold documents receive substantially higher DocScores than non-gold documents and are selected as Important Documents much more frequently. Under Llama, gold and non-gold documents receive mean DocScores of 0.647 and 0.324, respectively, with Important Document selection rates of 85.7% and 14.3%. Under Qwen, the same pattern holds: gold and non-gold documents receive mean DocScores of 0.722 and 0.348, with selection rates of 83.0% and 17.0%, respectively. This consistent separation across backbones suggests that DocScore captures document-level answerability rather than simply reflecting positional artifacts or model-specific scoring behavior.

Since DocScore is produced by an LLM-based Score Agent, its evidence selection could in principle reflect backbone-specific behavior rather than an intrinsic notion of evidence quality. To test this, we compare DocScore-based evidence selection across Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct under identical settings on Natural Questions, averaged over NQ-1/5/10/15/20. The cross-model selection results are reported in Table 12.

Our goal here is to measure cross-model *agreement*, not the absolute ability of either backbone to recover the gold document. Accordingly, when measuring gold evidence stability, we restrict the

denominator to cases where at least one backbone selects the gold document, and report how often both do. Cases where neither backbone selects the gold document are excluded, as they reflect retrieval difficulty rather than disagreement between backbones, and would otherwise conflate selection quality with cross-model consistency.

Disagreement is concentrated on distractors.

The selected document sets are not identical across backbones, but the disagreement is concentrated on distractors rather than gold evidence. When at least one backbone selects the gold document, both select it in 90.4% of cases. In contrast, among cases where both backbones select the gold document, their additionally selected non-gold documents differ 90.1% of the time.

Implication. High agreement on gold evidence alongside divergence on distractors shows that cross-model consistency is not an artifact of identical candidate pools: the two backbones converge on gold evidence while diverging on irrelevant context. DocScore therefore captures a stable answerability signal across the two evaluated backbones, supporting its use as a reliable quality signal for evidence pruning.

```

    "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 7: Prompt template for the Local and Global 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 8: Prompt template for the 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 9: Prompt template for the Score Agent.

Algorithm 1 DuQUAD Pipeline for a Single Question

Input: question Q , documents $\{\mathcal{D}_i\}_{i=1}^N$, local top- k , 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 $\mathcal{D}_i$

$z_i \leftarrow \frac{a_i - \mu_A}{\sigma_A} \quad \triangleright \text{z-score normalization } (\mu_A, \sigma_A: \text{mean/std of } \{a_i\}_{i=1}^N)$

$\Phi_{\text{cont},i} \leftarrow z_i - \frac{z_{i-1} + z_{i+1}}{2} \quad \triangleright \text{attention contrast } (i = 2, \dots, N-1)$

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

$\mathcal{I}_{\text{local}} \leftarrow \arg \text{top}_k \Phi_i \quad \triangleright \text{local document selection}$

$\mathcal{D}_{\text{local}} \leftarrow \{\mathcal{D}_i \mid i \in \mathcal{I}_{\text{local}}\}$

(2) Dual-view Agent Collaboration

$(\mathcal{H}_G, \mathcal{D}_G) \leftarrow \text{GlobalAgent}(Q, \{\mathcal{D}_i\}_{i=1}^N) \quad \triangleright \text{global reasoning + important docs}$

$(\mathcal{H}_L, \mathcal{D}_L) \leftarrow \text{LocalAgent}(Q, \mathcal{D}_{\text{local}}) \quad \triangleright \text{local reasoning + important docs}$

$\mathcal{D}_{\text{imp}} \leftarrow \mathcal{D}_G \cup \mathcal{D}_L \quad \triangleright \text{merge important docs}$

(3) DocScore-based Quality Gating

$(p_i, \tilde{a}_i) \leftarrow \text{ScoreAgent}(Q, \mathcal{D}_i), \quad \forall \mathcal{D}_i \in \mathcal{D}_{\text{imp}} \quad \triangleright \text{document reliability + answer span}$

$\mathcal{D}_{\text{keep}} \leftarrow \{\mathcal{D}_i \in \mathcal{D}_{\text{imp}} \mid p_i \geq \tau\} \quad \triangleright \text{DocScore-based pruning}$

(4) Evidence Snippet Extraction

$\mathcal{S}_i \leftarrow \text{Top2}_j \text{overlap}(u_{i,j}, \tilde{a}_i), \quad \forall \mathcal{D}_i \in \mathcal{D}_{\text{keep}} \quad \triangleright \text{select evidence sentences}$

$\mathcal{S} \leftarrow \bigcup_{\mathcal{D}_i \in \mathcal{D}_{\text{keep}}} \mathcal{S}_i \quad \triangleright \text{aggregate evidence}$

(5) Final Decision

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

return $\hat{y}$
