Mindscape-Aware Retrieval Augmented Generation for Improved Long Context Understanding

Anonymous ACL submission

Abstract

Humans understand long and complex texts by relying on a holistic semantic representation of the content. This global view helps organize prior knowledge, interpret new information, and integrate evidence dispersed across a document, as revealed by the *Mindscape-Aware Capability* of humans in psychology. However, current Retrieval-Augmented Generation (RAG) systems often lack explicit global semantic guidance, making it difficult to retrieve and integrate dispersed evidence in long-context tasks. In this paper, we propose Mindscape-Aware RAG (MiA-RAG), the first framework to formulate mindscape-aware retrieval and generation as a unified conditioning paradigm for LLM-based RAG. MiA-RAG builds a mindscape through hierarchical summarization and conditions both retrieval and generation on this global semantic representation. This enables the retriever to form enriched query embeddings and the generator to reason over retrieved evidence within a coherent global context. We evaluate MiA-RAG across diverse long-context and bilingual benchmarks for evidence-based understanding and global sense-making. It consistently surpasses baselines, and further analysis shows that it aligns local details with a coherent global representation, enabling more human-like long-context retrieval and reasoning.

1 Introduction

Human thinking is inherently context-dependent. For any learned topic, familiar situation, or ongoing project of engagement, humans maintain a global semantic representation in memory. When the same topic reappears, this global memory is reactivated, endowing humans with the **Mindscape-Aware capability** to become aware of the approximate scope of their knowledge and to rely on this memory to *interpret new inputs within context, selectively channel retrieval toward context-relevant knowledge, and guide subsequent reasoning accordingly*. This phenomenon is grounded in theories from psychology (Bartlett, 1932; Tulving and Thomson, 1973; Reyna and Brainerd, 1995) and neuroscience (Ralph et al., 2017), which posit that when a topic is reactivated, conscious cognition is constrained and guided by globally integrated knowledge, with converging support from neuroimaging observations. We review both the theoretical and empirical supports in Appx. A.

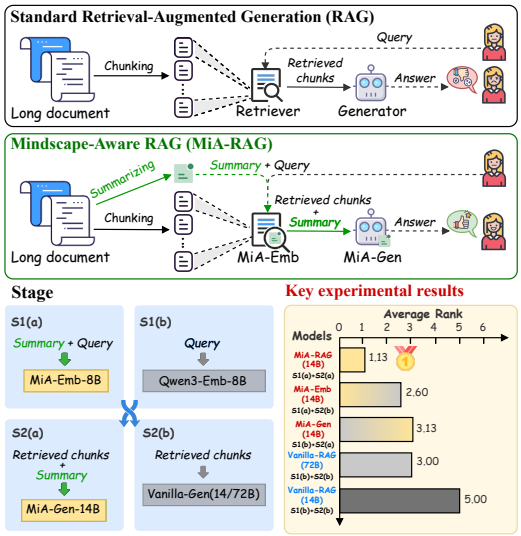

Figure 1: Average model ranks across five long-context benchmarks under 3/5/10-chunk settings.

Retrieval-Augmented Generation (RAG) (Zhu et al., 2025; Gao et al., 2023; Zhang et al., 2025a) has emerged as a critical strategy for long-context understanding by retrieving useful context fragments from very long inputs, thereby overcoming LLMs' limited context lengths (Lewis et al., 2020). However, current RAG systems primarily retrieve and generate based on local, evidence-level signals, lacking the *mindscape-aware* capability to activate a global semantic frame as humans do. Endowing RAG with this capability is therefore especially promising for personalized knowledge collec-

tions, such as long-context question answering (Bai et al., 2023), code generation (Wang et al., 2024b; Miao et al., 2024), and AI assistants over personal projects (Martin and Johnson, 2023). Specifically, the cognitive advantages of the mindscape translate naturally into the following benefits:

- ***Enriched Understanding***: supported by awareness of global semantics that fills missing information and resolves underspecified meanings.
- ***Selective Retrieval***: biases query embeddings toward the active topic's conceptual frame, filtering out ambiguities arising from other topics.
- ***Integrative Reasoning***: interprets retrieved results within the global context to ensure coherent synthesis and understanding.

Motivated by these insights, we propose the first approach to equip LLM-based RAG systems with mindscape-aware capabilities. Specifically, we tackle the long-context understanding problem by approximating the global impression of a long document with a summary generated in a hierarchical manner, which serves as an external representation of global memory. Taking this form of global memory as an additional input, we train models to fit two core functions of our new ***Mindscape-Aware RAG*** (**MiA-RAG**) framework:

- ***Mindscape-Aware Retrieval*** The mindscape-aware capability enables queries to be understood within their global semantic context, producing query representations that are not only anchored to the topical scope (*Selective Retrieval*), but also integrated with global contextual information (*Enriched Understanding*). We instantiate these capabilities through a specially trained embedding model, which takes both the global memory and the query as input, and learns to integrate global information into query embeddings to enhance retrieval performance.
- ***Mindscape-Aware Generation*** Solely enhancing the RAG pipeline with contextually informed query embeddings introduces a new challenge: the generator becomes weaker than the retriever, as it lacks access to the global context. Consequently, the generator may misinterpret the relevance of the retrieved information or fail to effectively utilize it, even when the retrieved content contains the correct evidence. To mitigate this asymmetry, we condition generation on the same global memory. By incorporating the global memory into the generator's inputs, the model learns to interpret the retrieved chunks and their relationship to the query within the broader global context (*Integrative Reasoning*), leading to more faithful reasoning and answers.

We evaluate MiA-RAG across a range of long-context understanding tasks spanning diverse domains and genres (*e.g.*, government reports, narratives), in both English and Chinese. The evaluation also covers various task formats, including free-form QA, multiple-choice QA, and claim verification, as well as different RAG configurations such as vanilla RAG and GraphRAG (Edge et al., 2024). As summarized in Figure 1, the MiA family is more effective than baselines; in particular, MiA-RAG-14B achieves the best average rank, surpassing the vanilla 72B system and highlighting the benefit of mindscape-aware retrieval and generation.

Beyond performance gains, we analyze MiA-RAG's internal mechanisms via embedding-space geometry and a new mindscape-coherent metric, showing that it internalizes global semantics: the mindscape reshapes query representations toward the global semantic space and acts as a scaffold that guides attention for Integrative Reasoning.

Our contributions are summarized as follows:

(1) We formulate the psycho- and neuro-inspired problem of mindscape-aware thinking, and present the first computational solution that equips LLMs with this capability.

(2) We conduct extensive experiments under diverse settings, demonstrating the necessity and effectiveness of integrating mindscape-aware capability into LLM-based systems.

(3) Our in-depth analysis reveals that the mindscape aligns the geometry of query representations with the global semantic space and serves as a semantic scaffold to guide attention, confirming the active internalization of global context rather than surface-level pattern matching.

2 Related Work

Context-Aware Embeddings Our MiA-Emb is related to the research topic of context-aware retrieval (Anthropic, 2024). This line of work mainly focuses on producing embeddings enriched with contextual information. A straightforward approach is to encode each chunk within a long-context window using LLMs that support extended inputs (Chen et al., 2024; Sturua et al., 2024; Nussbaum et al., 2024; Wang et al., 2024a; Lee et al., 2024; Li et al., 2023; Voyage-AI, 2025). To incorporate information beyond the local window, Xu et al. (2024) construct graphs of discourse relations

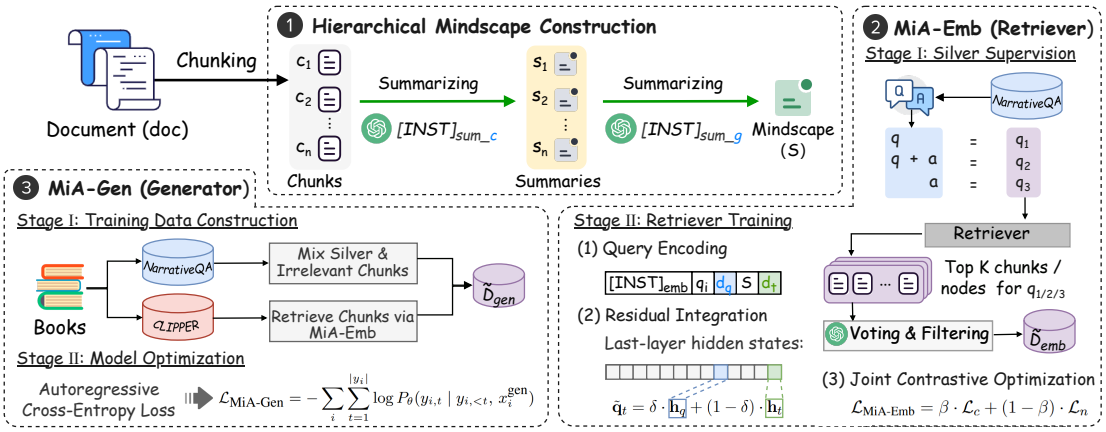

Figure 2: Overview of the framework of MiA-RAG.

and then utilize graph neighbors for augmentation. While these methods provide additional context, they do not directly teach the model how to fuse such information. Recently, (Wu et al., 2025) introduce training techniques that enable the embedding model to more effectively situate a chunk within its local neighborhood, achieving state-of-the-art retrieval performance with only modest training data requirements. However, it primarily enhances retrieval by enriching the chunk representation with local context. In contrast, our method injects global semantics into the query representation, guiding queries toward the correct semantic region of the index and improving selective retrieval without modifying the original chunk embeddings.

Long Context Compression To capture the overarching semantics of long documents, recent work moves beyond token retention or KV-cache compression (Yang et al., 2024; Li et al., 2024; Xiao et al., 2023) toward abstractive compression and global memory (Qian et al., 2025; Behrouz et al., 2024). MemoRAG (Qian et al., 2025) uses compressed memory for explicit planning, while LeanRAG (Zhang et al., 2026) organizes knowledge into hierarchical semantic graphs for structure-guided retrieval. In contrast, MiA-RAG learns to internalize the mindscape as a shared conditioning signal for both retrieval and generation, avoiding intermediate search steps or external graph construction.

3 Method

In this section, we introduce the Mindscape-Aware RAG (MiA-RAG) framework (Figure 2), which emulates how humans leverage global context to support understanding, retrieval, and reasoning.

3.1 Preliminaries

Let doc be a long document that has been partitioned into chunks $c_i \in C$. In a vanilla RAG pipeline, a retriever selects a set of chunks $C_{\text{ret}} \subseteq C$ for a query q, and the generator conditions on these chunks to produce an answer a:

$$q \xrightarrow{C_{\text{ret}}} a. \tag{1}$$

However, this setup does not provide a global view of the document for either retrieval or generation. To bridge this gap, we propose MiA-RAG, which incorporates an explicit global semantic scaffold termed the Mindscape S. By conditioning both retrieval and generation on S, MiA-RAG situates local evidence within a global context, improving retrieval accuracy and reasoning consistency:

$$q, S \xrightarrow{C_{ret}, S} a. \tag{2}$$

3.2 Hierarchical Mindscape Construction

We construct the Mindscape S via bottom-up hierarchical summarization.

We first prompt summarizer $\mathcal{M}_s$ (GPT-4o) with $\texttt{[INST]}_{\texttt{sum_c}}$ (Appx. F) to summarize each chunk:

$$s_i = \mathcal{M}_s(\texttt{[INST]}_{\texttt{sum_c}}, c_i). \tag{3}$$

After obtaining the chunk-level summaries, the sequence of $\{s_i\}$ is concatenated in order and further summarized using $\texttt{[INST]}_{\texttt{sum_g}}$ (Appx. F) to produce a single global representation:

$$S = \mathcal{M}_s(\texttt{[INST]}_{\texttt{sum_g}}, [s_1, s_2, \ldots, s_n]). \tag{4}$$

The resulting S serves as a document-level summary and can be updated incrementally by regenerating only the affected paths from modified chunks to the root (Chang et al., 2023).

3.3 MiA-Emb: Mindscape-Aware Retriever

To obtain the Mindscape-Aware Retriever (MiA-Emb), we construct a supervision dataset and optimize the model with a multi-task objective.

3.3.1 Supervision Construction

Existing long-narrative understanding datasets such as NarrativeQA (Kočiskỳ et al., 2018) typically provide only QA pairs and do not link questions to fine-grained supporting evidence. Such supervision is essential for training long-context retrievers, whether evidence is represented as text chunks in standard RAG or as graph nodes in GraphRAG.

Given the cost of manual annotation, we automatically extend NarrativeQA to curate $\tilde{D}_{\text{emb}}$, a dataset offering silver-standard alignments at both chunk and node levels. For chunk evidence, we perform *query augmentation*, *majority-vote ensemble* retrieval, and *LLM-based filtering* to identify silver chunks (Algorithm 1 in Appx. B). As shown in Table 1, an oracle experiment validates these annotations: using only silver chunks ($\sim$1/30 tokens) already exceeds full-context performance. Node evidence is constructed in an analogous way.

In total, $\tilde{D}_{\text{emb}}$ comprises 27,117 questions, averaging 2.3 silver chunks and 2.9 nodes per question. Details of the construction are provided in Appx. B.

Input	# Books	Avg. Chunks	EM	F1
Silver Chunks	20	2.3	**31.29**	**52.91**
Oracle (Full Book)	20	69.2	30.04	50.79

Table 1: Validation of silver chunk annotations.

3.3.2 Model Optimization

We develop MiA-Emb to explicitly inject global context into query representations. Specifically, we encode each query q_i by explicitly incorporating the summary S together with task-specific control tokens d. The input sequence is defined as

$$Q = [\texttt{[INST]}_{\texttt{emb}};\ q_i;\ d_q;\ S;\ d_t], \tag{5}$$

where $\texttt{[INST]}_{\texttt{emb}}$ (Appx. F) is the instruction, d_q marks the end of query, and $d_t = [d_n, d_c]$ activates node- and chunk-retrieval modes, respectively.

To balance the query intent with global guidance, we use a residual integration and train the model using an InfoNCE objective (van den Oord et al., 2018). Full details are provided in Appx. B.3.

3.4 MiA-Gen: Mindscape-Aware Generator

3.4.1 Training Data Construction

We develop the Mindscape-Aware Generator (**MiA-Gen**) by fine-tuning on a composite corpus derived from the training splits of two datasets: **NarrativeQA** for long context QA and **CLIPPER** (Pham et al., 2025), a synthetic dataset featuring narrative claims for verification. Each training instance is formatted as follows:

$$\underbrace{\langle \texttt{[INST]}_{\texttt{gen}}\ ;\ S;\ \hat{C}_{\text{ret}};\ q_i \rangle}_{x_i^{\text{gen}}} \rightarrow y_i, \tag{6}$$

where $\hat{C}_{\text{ret}}$ is the retrieved chunks, and y_i is the target. Instruction $\texttt{[INST]}_{\texttt{gen}}$ is detailed in Appx. F.

To ensure robustness against retrieval noise, we construct $\hat{C}_{\text{ret}}$ to simulate realistic inference conditions. For NarrativeQA, we mix silver evidence chunks with irrelevant text; for CLIPPER, we utilize the actual top-k retrieval results from MiA-Emb. Furthermore, we augment the data by varying the number of chunks in $\hat{C}_{\text{ret}}$ ($k \in \{3, 5, 10\}$) for each query, enabling the model to handle diverse context lengths. The aggregation of these instances constitutes the final dataset for the generator $\tilde{\mathcal{D}}_{\text{gen}}$.

3.4.2 Model Optimization

MiA-Gen is optimized over $\tilde{\mathcal{D}}_{\text{gen}}$ using the autoregressive cross-entropy loss:

$$\mathcal{L}_{\text{MiA-Gen}} = -\sum_{i} \sum_{t=1}^{|y_i|} \log P_\theta(y_{i,t} \mid y_{i,<t},\ x_i^{\text{gen}}). \tag{7}$$

4 Experimental Setting

Evaluated Models MiA-RAG is implemented by fine-tuning Qwen-series (Zhang et al., 2025b; Team, 2024) models. We develop two components:

- **Mindscape-Aware Retriever**: we fine-tune *Qwen3-Embedding-8B* to obtain MiA-Emb.
- **Mindscape-Aware Generator**: we fully fine-tune *Qwen2.5-14B-Instruct* to obtain MiA-Gen.

We compare against Qwen baselines and state-of-the-art RAG methods, including HippoRAG v2 (Guti'errez et al., 2025), MemoRAG (Qian et al., 2025) and RAPTOR (Sarthi et al., 2024).

Public Long Narrative Understanding Tasks We evaluate on long-narrative understanding benchmarks, many of which exceed common LLM input limits (e.g., 128K tokens). Since our model is trained on **NarrativeQA** (Kočiskỳ et al., 2018), we conduct in-domain evaluation on its held-out books, and perform out-of-domain evaluation on the EN.MC subset of ∞**Bench** (Zhang et al., 2024), the public subset of **Nocha** (Karpinska et al., 2024), and **DetectiveQA-En/Zh** (Xu et al., 2025).

Dataset statistics, baselines, experimental details, and computational cost are provided in Appx. C.

Model	Retriever		Generator		NarrativeQA	∞ Bench	Det.QA-Zh	Det.QA-En	NoCha	Avg.
	Emb. Model	+S	Gen. Model	+S	F1	Acc	Acc	Acc	Pair Acc	
Summary-Only	-	-	Qwen2.5-72B	✓	39.24	72.05	73.67	61.33	31.75	55.61
Vanilla-RAG	Qwen3-Emb-8B	✗	Qwen2.5-72B	✗	41.13/45.51/49.06	75.55/80.79/86.90	63.67/70.83/78.00	55.50/61.33/71.17	33.33/38.10/41.27	59.48
HippoRAG-v2	Qwen3-Emb-8B	✗	Qwen2.5-72B	✗	40.06/45.20/47.16	79.48/84.28/85.15	70.33/73.83/77.17	64.83/68.00/68.50	33.33/36.51/47.62	61.43
MemoRAG	BGE-M3	✗	Qwen2.5-72B	✗	40.72/43.63/44.86	75.11/76.42/78.17	67.50/68.83/71.17	61.17/63.50/64.33	**46.03/49.21**/49.21	59.06
Raptor	Qwen3-Emb-8B	✗	Qwen2.5-72B	✗	39.48/42.76/46.94	75.98/79.91/83.84	62.00/68.67/74.67	54.50/62.00/69.67	34.92/34.92/46.03	58.98
MiA (Gen-Only)	Qwen3-Emb-8B	✗	Qwen2.5-72B	✓	47.67/48.46/51.81	82.10/83.84/86.46	76.00/78.50/81.33	68.17/69.67/73.33	36.51/42.86/44.44	64.74
MiA (Emb-Only)	MiA-Emb-8B	✓	Qwen2.5-72B	✗	46.38/48.06/49.88	**84.72/87.77/90.39**	76.17/81.17/82.67	67.17/71.83/75.33	42.86/42.86/49.21	66.43
MiA	MiA-Emb-8B	✓	Qwen2.5-72B	✓	**50.05/51.04/53.15**	84.71/86.46/88.21	**81.67/83.17/84.17**	**70.33/72.33/75.50**	41.27/44.44/**52.38**	**67.93**
Summary-Only	-	-	Qwen2.5-14B	✓	38.03	61.57	70.50	58.00	17.46	49.11
Vanilla-RAG	Qwen3-Emb-8B	✗	Qwen2.5-14B	✗	39.32/40.99/44.38	72.49/73.80/77.29	60.33/68.50/75.67	55.17/59.83/66.67	17.46/11.11/15.87	51.93
HippoRAG-v2	Qwen3-Emb-8B	✗	Qwen2.5-14B	✗	38.34/41.28/43.60	76.86/79.04/80.79	65.67/71.00/75.00	58.00/61.33/65.83	20.63/15.87/17.46	54.05
MemoRAG	BGE-M3	✗	Qwen2.5-14B	✗	36.58/42.87/44.42	71.18/72.05/73.80	60.70/66.17/68.83	54.33/55.67/58.50	28.57/31.75/34.92	53.35
Raptor	Qwen3-Emb-8B	✗	Qwen2.5-14B	✗	36.81/39.92/44.00	75.55/82.10/82.53	60.50/63.83/71.33	53.50/60.00/65.50	22.22/14.29/15.87	54.27
MiA (Gen-Only)	Qwen3-Emb-8B	✗	Qwen2.5-14B	✓	43.04/43.32/45.83	75.98/80.79/79.48	75.33/78.83/80.17	65.33/66.83/70.00	19.05/15.87/26.98	57.79
MiA (Gen-Only)	Qwen3-Emb-8B	✗	MiA-Gen-14B	✓	50.55/50.08/51.99	75.98/82.10/80.79	76.17/78.50/79.50	67.67/**71.50**/71.67	49.21/47.62/50.79	65.61
MiA (Emb-Only)	MiA-Emb-8B	✓	Qwen2.5-14B	✗	45.89/46.77/47.13	79.48/82.97/84.28	73.00/77.17/80.83	62.33/65.83/70.33	22.22/26.98/26.98	59.48
MiA	MiA-Emb-8B	✓	Qwen2.5-14B	✓	44.38/46.66/47.87	79.91/**83.41**/84.28	78.33/79.00/**82.00**	66.50/69.17/71.82	30.16/28.57/38.10	62.01
MiA-RAG	MiA-Emb-8B	✓	MiA-Gen-14B	✓	**52.48/53.52/53.56**	**80.79**/81.22/**85.15**	**79.00/79.67**/81.83	**69.00**/71.17/**75.50**	**55.56/49.21/53.97**	**68.11**

Table 2: Overall results on Long-story QA and Reasoning benchmarks under top-3/5/10 retrieval. "+S" indicates whether the summary is incorporated at that stage. Darker gray rows represent deeper mindscape involvement; bold marks the best per scale. The final column reports the average over all metrics per row. Notably, MemoRAG (Qian et al., 2025) uses an additional 7B memory model to enrich retrieval, as detailed in Appx. C.2.

5 Experiments

5.1 Study I: Retrieval Results

Table 3 summarizes retrieval performance. MiA-Emb consistently outperforms all baselines across the benchmarks, and even surpasses **Sit-Emb** (Wu et al., 2025), a state-of-the-art model specialized for story understanding. Further comparisons with other embedding models are given in Appx. D.1.

MiA-Emb achieves strong gains on in-domain NarrativeQA and transfers well to out-of-domain DetectiveQA, demonstrating its ability to leverage global context for accurate cross-domain evidence localization. We further validate this generalization on the long-term dialogue benchmark LoCoMo (Maharana et al., 2024) in Appx. D.2.

Finally, our ablation study shows that removing the summary (*w/o Summary*) leads to substantial degradation, underscoring the role of the mindscape representation in enhancing retrieval.

Method	NarrativeQA			DetectiveQA-ZH			DetectiveQA-EN		
	3	5	10	3	5	10	3	5	10
MiA-Emb-8B	**62.68**	**75.92**	**88.09**	**46.75**	**59.17**	**72.50**	**42.08**	**54.17**	**69.75**
w/o Summary	55.62	67.19	83.65	37.92	48.75	66.50	34.00	45.75	61.25
Sit-Emb-8B	59.98	70.70	82.68	42.50	54.50	69.30	36.75	49.25	63.83
Qwen-Emb-8B	41.81	54.51	71.13	28.58	39.08	55.58	24.17	34.17	49.25

Table 3: Retrieval performance measured by Recall(%).

5.2 Study II: Long Narrative Understanding

5.2.1 End-Task Results

Table 2 presents the complete RAG pipeline evaluation across five long-context benchmarks. We also include results on the Helmet (Yen et al., 2024) version of NarrativeQA to compare with long-window LLM solutions in Appx. D.3, on which our MiA-RAG improves over GPT-4o with only ∼3% of input tokens. There are three key takeaways.

MiA-RAG Achieves Superior Overall Performance. MiA-RAG attains the best results across all benchmarks, spanning English and Chinese, diverse domains, and multiple task formats, with gains of +16.18% over vanilla-RAG-14B and +8.63% over vanilla-RAG-72B, showing the effectiveness of mindscape capability for RAG systems.

Mindscape-Aware Retrieval Consistently Improves Performance As verified in Sec. 5.1, MiA-Emb substantially enhances retrieval quality. When integrated into the full pipeline, substituting the vanilla retriever with MiA-Emb yields consistent gains. MiA-Emb improves the average scores of the 72B and 14B generators by 6.95% and 7.55%, respectively, confirming that globally informed queries yield more effective retrieval.

Integrative Reasoning Benefits from Mindscape-Conditioned Generation Simply supplying the summary to a vanilla generator yields a consistent **+3.79%** improvement, showing that global contextual cues provide useful guidance. A larger gain is observed when the generator is fine-tuned under the same mindscape-conditioning paradigm as the retriever. Under identical inputs, our MiA-Gen-14B achieves a substantially larger **+11.16%** gain. This disparity suggests that MiA-Gen more effectively integrates retrieved chunks with the global semantics that guided their selection.

5.2.2 Ablation Study

We perform ablations to assess each MiA-RAG component, with results shown in Tables 3 and 4.

Method	Narra.QA	∞Bench	DetectiveQA	NoCha
MiA-Gen-14B	**53.19**	82.39	**76.03**	**52.91**
w/o Summary	50.49	75.39	71.03	44.97
w/o Claim.	51.22	**85.44**	75.58	44.44
w/o QA	46.40	81.08	72.81	46.56

Table 4: Ablation study of MiA-Gen-14B. Reported scores are averaged over the 3/5/10-chunk settings.

Summary Generator	Recall@3/5/10 (%)	F1-Score (%)
GPT-4o (Ours)	62.68/75.92/88.09	52.48/53.52/53.56
Qwen2.5-32B-Instruct	61.66/74.60/88.06	50.20/51.80/53.37
Qwen2.5-14B-Instruct	59.74/73.54/87.68	51.45/51.81/52.61
Qwen2.5-7B-Instruct	58.62/72.61/86.17	49.79/51.55/51.48

Table 5: Impact of summary quality on NarrativeQA.

Impact of Mindscape-Conditioning Same as in the embedding stage (Table 3), removing the summary (*w/o Summary*) leads to substantial degradation in the generation stage (Table 4). These declines indicate that summary-based supervision helps align queries with global semantics and supports the integration of dispersed evidence.

Benefit of Multi-Paradigm Supervision Ablating either supervision paradigm (*w/o Claim.* for claim verification or *w/o QA* for question answering) consistently degrades performance, indicating that exposure to diverse reasoning patterns improves generalization beyond any single task.

5.3 Study III: MiA-GraphRAG for Global QA

We further evaluate MiA-RAG for global sense-making in a GraphRAG QA setting, where it retrieves relevant graph nodes (entities) for holistic document understanding and achieves clear gains over baselines. Details are reported in Appx. D.4.

5.4 Study IV: Impact of Model Scales

We evaluate the scalability of MiA-Emb across backbone sizes (0.6B to 8B) against `SFT-Emb` (identical to the *w/o Summary* ablation in Sec. 5.1, i.e., trained and evaluated without summaries) and `Vanilla` Qwen3-Embedding baselines. As shown in Figures 3 and 9, MiA-Emb consistently outperforms both baselines; notably, MiA-Emb-0.6B already surpasses the Vanilla 8B model in both retrieval recall and downstream QA and reasoning performance. We also scale MiA-Gen across model sizes and observe consistent gains over the vanilla Qwen2.5-Instruct models (1.5B∼72B), presented in Figure 4. In particular, **MiA-Gen-14B** even surpasses the 72B model. These results demonstrate that our approach exhibits scalability across model sizes. Numerical results are provided in Appx. D.5.

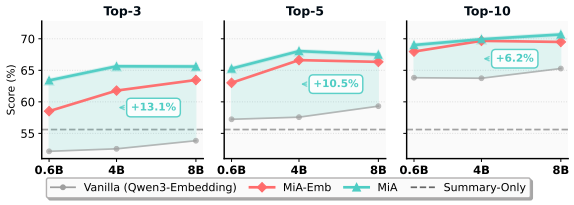

Figure 3: Impact of retriever scale on average results across benchmarks (DetectiveQA-ZH/EN,∞Bench, NoCha, NarrativeQA) with a Qwen2.5-72B generator.

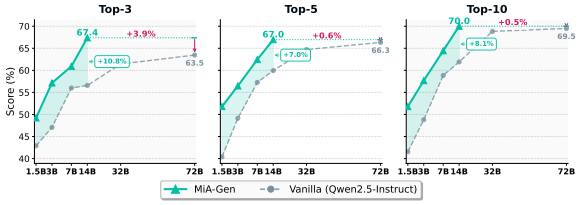

Figure 4: Impact of generator model scale on average results over 5 benchmarks, with a MiA-Emb-8B retriever.

5.5 Study V: Impact of Summary Quality

We examine the robustness of MiA-RAG to summary quality by replacing GPT-4o summaries with those generated by open-source Qwen2.5 models (7B, 14B, and 32B). As shown in Table 5, MiA-RAG maintains stable performance across summarizers, with the 14B model achieving results comparable to GPT-4o. It is also robust to summary length: NarrativeQA summaries range from 240 to 1,243 tokens but show only a weak correlation with Recall@10 (Pearson $r = 0.23$). These results indicate that summary information loss has a limited impact, since the mindscape serves as a global scaffold for retrieval and reasoning rather than a lossless store of fine-grained evidence. We further validate this observation in Sec. 6.1 and Sec. 6.2.

6 Analysis

In this section, we introduce analytical methods to evaluate whether the resulting MiA-RAG exhibits the three hypothesized capabilities proposed in the introduction, namely *Enriched Understanding*, *Selective Retrieval*, and *Integrative Reasoning*.

6.1 The Role of Global Summaries

While ablation confirms that MiA-RAG benefits from incorporating summaries, a key question remains: what is their functional role during inference? We first show that summaries are not useful simply because they cover the answer. We evalu-

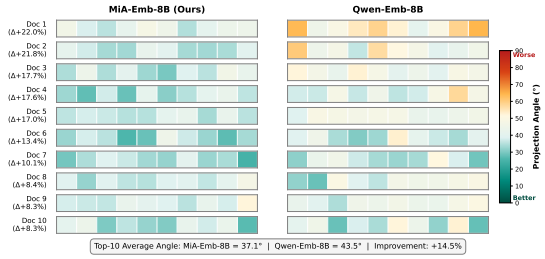

Figure 5: Comparison of projection angles for MiA-Emb and Qwen3-Emb. Lower angles indicate better alignment of queries with the book's semantic subspace.

ate a *Summary-Only* variant where the generator predicts answers using only the summary (Table 2). This variant consistently underperforms vanilla-RAG and falls far short of MiA-RAG. These results, together with same-summary retrieval controls (Table 14), indicate that summaries function not as standalone evidence or auxiliary text, but as *semantic scaffolds* for retrieval and reasoning.

6.2 Geometric Properties of the MiA Embedding Space

Extending the Sec. 6.1, we further examine:

(*H1*) *MiA-Emb facilitates Selective Retrieval.*

That is, whether the embedding model biases query representations toward the active book topic, thereby better positioning them within the subspace supported by the corresponding chunks.

Method We visualize query and chunk embeddings with t-SNE (Maaten and Hinton, 2008). To characterize the semantic structure of the document, We first fit t-SNE on the chunk embeddings only, yielding a 2D manifold that reflects the document's semantic structure. We then embed the query representations into the same 2D space and inspect how well each model positions queries relative to the corresponding topic-relevant chunk regions.

Results Figure 5 shows a clear geometric distinction between MiA-Emb and the vanilla embedding model. Across books, MiA-Emb consistently yields smaller projection angles, meaning that query embeddings lie closer to the semantic subspaces spanned by their corresponding documents. On average, MiA-Emb-8B achieves 37.1°, compared with 43.5° for Qwen-Emb-8B, demonstrating that mindscape conditioning more effectively guides queries toward the correct semantic region and enables more precise selective retrieval.

6.3 Residual Stream and Attention Analysis of the MiA Embedding Model

We analyze the model's internal representations and attention to examine the following hypothesis:

(*H2*) *MiA-Emb facilitates* ***Enriched Understanding*** *of queries.*

We verify the hypothesis in two folds. First, we examine whether performance gains from MiA-Emb correlate with increased use of the global summary. If so, we then study whether the model focuses its attention on information that can enrich the queries in these situations.

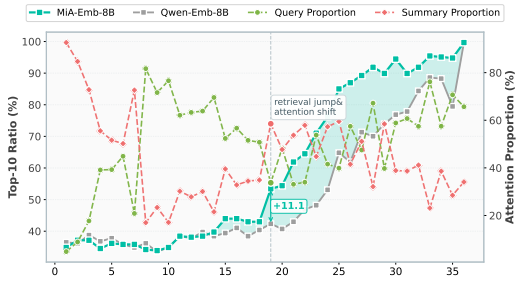

Figure 6: Layer-wise comparison of silver-chunk retrieval accuracy and attention allocation proportion.

(*H2.1*) *MiA-Emb puts increased attention to the global summary at layers of improved predictability compared to the baseline.*

Method Following the approach of (Jiang et al., 2024), we compare MiA-Emb with the vanilla embedding model through their residual streams to analyze how retrieval-relevant information is progressively accumulated into the query representation. To ensure comparability, we select 100 queries for which both models achieve Recall@10 = 100%. Concretely, we track the layer-wise Top-10 silver-chunk ratio for both models, reflecting how the hidden states at each layer steer the retrieval distribution toward the correct evidence. For MiA-Emb, we additionally examine the attention from the last token to summary tokens and to query tokens, enabling us to assess whether improvements coincide with increased use of global-summary cues.

Results As shown in Figure 6, MiA-Emb exhibits a clear rise in silver-chunk recall beginning at the middle layers. This rise coincides with increased attention to the global summary in the same layer range, suggesting that the model progressively injects summary-derived cues into the query representation. This incorporation of global signals enriches the query embedding, enabling MiA-Emb

Query: Where do Conan and Olivia seek refuge to take a rest?"

Answer: In ancient ruins on a deserted island.

Summary:
Olivia, a fugitive princess sold into slavery,.... After Conan kills **Shah** Amurath, Olivia pleads for his protection, and together they escape across the **inland** sea Vilayet in a small boat.\n\nDuring their journey, **Olivia** reveals her tragic past while Conan recounts his recent life as a mercenary and fugitive. They reach a seemingly **deserted island**, **where** they encounter unsettling oddities: ..., and ancient **ruins** filled with eerie, lifelike iron statues with hawk-like faces. The **statues**' uncanny realism disturbs them, but as night falls, **they** reluctantly **seek shelter in the ruins**.That night, **Olivia** dreams of an ....

Figure 7: Attention pattern of MiA-Emb: the last token attends to preceding summary tokens, with red regions indicating tokens that receive high attention.

to develop a deeper semantic understanding of the query and thus support more selective retrieval.

(*H2.2*) *MiA-Emb attends to information that enriches the query at the layers identified in* (H2.1).

Method To understand how summary information enriches query representations, we inspect the summary-attentive layers identified in Section (*H2.1*). Our goal is to assess whether the embedding token allocates its attention to summary tokens that are semantically aligned with the query. If such attention emerges precisely at layers where retrieval performance improves, it suggests that MiA-Emb enhances query understanding through targeted integration of global context.

Results Figure 7 shows that, at the layers corresponding to retrieval gains, the final embedding token concentrates its attention on summary phrases that correspond to answer-relevant entities, events, or locations. This indicates that MiA-Emb selectively extracts semantically aligned global cues and integrates them into the query representation, reinforcing the layer-wise analysis and confirming that MiA-Emb enhances query understanding via summary-based enrichment.

6.4 Attention Pattern Analysis in the Generation Model

(*H3*) *MiA-Gen facilitates* ***Integrative Reasoning*** *over retrieved chunks within the global mindscape.*

To examine how the mindscape steers generation toward relevant evidence, we introduce the Mindscape-Coherent Evidence Alignment metric (MCEA). MCEA measures whether chunks receiving stronger mindscape attention are also preferentially attended by the query, comparing this coupling between relevant and noise chunks. Thus,

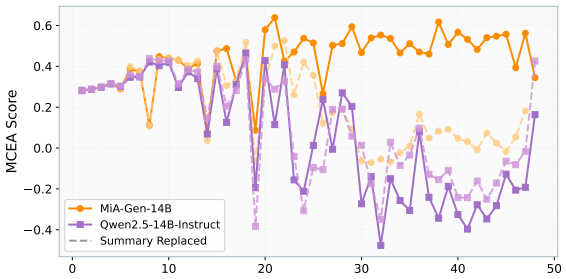

Figure 8: Layer-wise Mindscape-Coherent Evidence Alignment (MCEA) scores for generator.

MCEA serves as an attention-level diagnostic of mindscape-conditioned evidence integration, rather than a raw score for direct cross-model comparison. A formal definition is given in Appx. E.

Method We compute MCEA for MiA-Gen and Qwen2.5-14B-Instruct at each transformer layer. To test whether the observed coupling depends on valid global information rather than incidental attention, or positional bias, we introduce a *summary-replaced* control that replaces the original summary with an unrelated text of the same length.

Results Figure 8 highlights two key findings. First, MiA-Gen exhibits consistently high MCEA values, especially in middle and late layers. This suggests that local chunks first absorb global mindscape information, after which the query increasingly attends to the enriched evidence. Second, replacing the valid summary with unrelated text sharply reduces MiA-Gen's MCEA, showing that its evidence integration depends on valid mindscape. In contrast, the vanilla model shows little sensitivity to this perturbation; its raw MCEA under unrelated summaries can remain comparable, likely reflecting incidental attention coupling rather than meaningful mindscape-to-evidence transfer.

7 Conclusion

Inspired by human cognitive ability to interpret inputs within a global mindscape, we propose MiA-RAG, a mindscape-aware framework for LLM-based RAG. MiA-RAG approximates this global impression with hierarchical summaries as persistent global memory, conditioning both retrieval and generation on it. MiA-RAG achieves strong performance on evidence-based long-context understanding and global sense-making tasks. Empirical analysis shows that the mindscape supports global semantic query understanding, selective retrieval, and integrative reasoning over dispersed evidence.

Limitations

While MiA-RAG demonstrates strong performance on narrative long-context QA and reasoning, it relies on a precomputed global summary as the mindscape representation, which introduces additional preprocessing overhead compared with standard RAG. Although this cost is incurred only once per document and can be amortized across subsequent queries, it remains a practical consideration for deployment. At the same time, the hierarchical bottom-up design supports incremental updates by regenerating only the affected paths, which partially mitigates this limitation in dynamic or evolving document settings. In addition, our experiments mainly focus on narrative long-context understanding, while other application settings remain unexplored. Nevertheless, the strong performance on a global sense-making dataset provides initial evidence that MiA-RAG may generalize beyond purely narrative QA settings. Finally, part of the supervision signal is derived from commercial LLMs, which may introduce latent bias or hallucinated content. Nevertheless, the empirical results suggest that the mindscape-aware training strategy remains robust under imperfect supervision.

References

Anthropic. 2024. Introducing contextual retrieval.

Sam Audrain and Mary Pat McAndrews. 2022. Schemas provide a scaffold for neocortical integration of new memories over time. *Nature communications*, 13(1):5795.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, and 1 others. 2023. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*.

Frederic Charles Bartlett. 1932. *Remembering: A study in experimental and social psychology*. Cambridge university press.

Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. 2024. Titans: Learning to memorize at test time. *arXiv preprint arXiv:2501.00663*.

Jeffrey R Binder, Rutvik H Desai, William W Graves, and Lisa L Conant. 2009. Where is the semantic system? a critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral cortex*, 19(12):2767–2796.

Garvin Brod, Ulman Lindenberger, and Yee Lee Shing. 2017. Neural activation patterns during retrieval of schema-related memories: Differences and commonalities between children and adults. *Developmental science*, 20(6):e12475.

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2023. Booookscore: A systematic exploration of book-length summarization in the era of llms. *arXiv preprint arXiv:2310.00785*.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

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).

Samuel J Gershman, Anna C Schapiro, Almut Hupbach, and Kenneth A Norman. 2013. Neural context reinstatement predicts memory misattribution. *Journal of Neuroscience*, 33(20):8590–8595.

Asaf Gilboa and Hannah Marlatte. 2017. Neurobiology of schemas and schema-mediated memory. *Trends in cognitive sciences*, 21(8):618–631.

Bernal Jim'enez Guti'errez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. 2025. From rag to memory: Non-parametric continual learning for large language models. In *arXiv.org*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Che Jiang, Biqing Qi, Xiangyu Hong, Dayuan Fu, Yang Cheng, Fandong Meng, Mo Yu, Bowen Zhou, and Jie Zhou. 2024. On large language models' hallucination with regard to known facts. 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 1041–1053.

Marzena Karpinska, Katherine Thai, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. One thousand and one pairs: A "novel" challenge for long-context language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*.

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

James E Kragel, Youssef Ezzyat, Bradley C Lega, Michael R Sperling, Gregory A Worrell, Robert E Gross, Barbara C Jobst, Sameer A Sheth, Kareem A Zaghloul, Joel M Stein, and 1 others. 2021. Distinct cortical systems reinstate the content and context of episodic memories. *Nature Communications*, 12(1):4444.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2025. Nv-embed: Improved techniques for training llms as generalist embedding models. *Preprint*, arXiv:2405.17428.

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.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. 2024. Snapkv: Llm knows what you are looking for before generation. *Advances in Neural Information Processing Systems*, 37:22947–22970.

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*.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 13851–13870. Association for Computational Linguistics.

Raiza Martin and Steven Johnson. 2023. Introducing notebooklm.

Jing Miao, Charat Thongprayoon, Supawadee Suppadungsuk, Oscar A Garcia Valencia, and Wisit Cheungpasitporn. 2024. Integrating retrieval-augmented generation with large language models in nephrology: advancing practical applications. *Medicina*, 60(3):445.

Zach Nussbaum, John X Morris, Brandon Duderstadt, and Andriy Mulyar. 2024. Nomic embed: Training a reproducible long context text embedder. *arXiv preprint arXiv:2402.01613*.

Chau Minh Pham, Yapei Chang, and Mohit Iyyer. 2025. Clipper: Compression enables long-context synthetic data generation. *arXiv preprint arXiv:2502.14854*.

Hongjin Qian, Zheng Liu, Peitian Zhang, Kelong Mao, Defu Lian, Zhicheng Dou, and Tiejun Huang. 2025. Memorag: Boosting long context processing with global memory-enhanced retrieval augmentation. In *Proceedings of the ACM on Web Conference 2025*, pages 2366–2377.

Matthew A Lambon Ralph, Elizabeth Jefferies, Karalyn Patterson, and Timothy T Rogers. 2017. The neural and computational bases of semantic cognition. *Nature reviews neuroscience*, 18(1):42–55.

Valerie F Reyna and Charles J Brainerd. 1995. Fuzzy-trace theory: An interim synthesis. *Learning and individual Differences*, 7(1):1–75.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher Manning. 2024. Raptor: Recursive abstractive processing for tree-organized retrieval. In *International Conference on Learning Representations*, volume 2024, pages 32628–32649.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Andreas Koukounas, Nan Wang, and Han Xiao. 2024. jina-embeddings-v3: Multilingual embeddings with task lora. *Preprint*, arXiv:2409.10173.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Endel Tulving and Donald M Thomson. 1973. Encoding specificity and retrieval processes in episodic memory. *Psychological review*, 80(5):352.

Aäron van den Oord, Yazhe Li, and O. Vinyals. 2018. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748.

Voyage-AI. 2025. Introducing voyage-context-3: focused chunk-level details with global document context. Blog post.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916.

Zora Zhiruo Wang, Akari Asai, Xinyan Velocity Yu, Frank F Xu, Yiqing Xie, Graham Neubig, and Daniel Fried. 2024b. Coderag-bench: Can retrieval augment code generation? *arXiv preprint arXiv:2406.14497*.

Junjie Wu, Jiangnan Li, Yuqing Li, Lemao Liu, Liyan Xu, Jiwei Li, Dit-Yan Yeung, Jie Zhou, and Mo Yu. 2025. Sitemb-v1. 5: Improved context-aware dense retrieval for semantic association and long story comprehension. *arXiv preprint arXiv:2508.01959*.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2023. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*.

Liyan Xu, Jiangnan Li, Mo Yu, and Jie Zhou. 2024. Fine-grained modeling of narrative context: A coherence perspective via retrospective questions. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5822–5838.

Yangwen Xu, Qixiang Lin, Zaizhu Han, Yong He, and Yanchao Bi. 2016. Intrinsic functional network architecture of human semantic processing: Modules and hubs. *Neuroimage*, 132:542–555.

Zhe Xu, Jiasheng Ye, Xiaoran Liu, Xiangyang Liu, Tianxiang Sun, Zhigeng Liu, Qipeng Guo, Linlin Li, Qun Liu, Xuanjing Huang, and Xipeng Qiu. 2025. DetectiveQA: Evaluating long-context reasoning on detective novels. In *Workshop on Reasoning and Planning for Large Language Models*.

Dongjie Yang, XiaoDong Han, Yan Gao, Yao Hu, Shilin Zhang, and Hai Zhao. 2024. Pyramidinfer: Pyramid kv cache compression for high-throughput llm inference. *arXiv preprint arXiv:2405.12532*.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. 2024. Helmet: How to evaluate long-context language models effectively and thoroughly. *arXiv preprint arXiv:2410.02694*.

Qinggang Zhang, Shengyuan Chen, Yuanchen Bei, Zheng Yuan, Huachi Zhou, Zijin Hong, Hao Chen, Yilin Xiao, Chuang Zhou, Junnan Dong, and 1 others. 2025a. A survey of graph retrieval-augmented generation for customized large language models. *arXiv preprint arXiv:2501.13958*.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Khai Hao, Xu Han, Zhen Leng Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. 2024. ∞bench: Extending long context evaluation beyond 100k tokens. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, and 1 others. 2025b. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Yaoze Zhang, Rong Wu, Pinlong Cai, Xiaoman Wang, Guohang Yan, Song Mao, Ding Wang, and Botian Shi. 2026. Leanrag: Knowledge-graph-based generation with semantic aggregation and hierarchical retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 34862–34869.

Zulun Zhu, Tiancheng Huang, Kai Wang, Junda Ye, Xinghe Chen, and Siqiang Luo. 2025. Graph-based approaches and functionalities in retrieval-augmented generation: A comprehensive survey.

A Supports of Mindscape-Aware Capabilities in Broader Research Fields

We show that the existence and advantages of mindscape-aware capabilities are supported by research on human memory in psychology and neuroscience.

Supports in Psychology The existence of mindscape-aware capability traces back to the concept of schema (Bartlett, 1932) and aligns with Fuzzy-Trace Theory (FTT; Reyna and Brainerd 1995). Schemas provide integrated structures for familiar topics, guiding attention and constraining interpretation during information processing. FTT further posits that human memory encodes experiences at two complementary levels: verbatim traces that preserve surface details and gist traces that capture abstract, meaning-based structure. When new information relates to a familiar topic, gist-level representations are reactivated, providing a global semantic scaffold for interpretation, retrieval, and reasoning. Our mindscape-aware framework can be viewed as a computational approximation of such gist-based cognition in retrieval-augmented reasoning systems.

The advantages of mindscape-aware capability are also related to the Encoding Specificity Principle (Tulving and Thomson, 1973). This principle states that reinstating the original contextual pattern of a familiar topic or task can reactivate the corresponding memory network, thereby improving retrieval effectiveness and interpretive coherence.

Supports in Neuroscience Mindscape-aware capabilities are also supported by neuroscience research. The *controlled semantic cognition (CSC)* framework (Ralph et al., 2017) suggests that cognition is guided by globally integrated semantic

Algorithm 1 Silver Evidence Annotation

Input: Dataset $\mathcal{D} = \{(q_i, a_i)\}_{i=1}^{N}$, mindscape summary S, retriever $\mathcal{E}_s$, task $t \in \{c, n\}$.
Define: Evidence units $U = C$ (if $t = c$) or $U = V$ (if $t = n$).
Output: Silver-annotated dataset $\tilde{\mathcal{D}}^t_{\text{emb}} = \{(q_i, \tilde{U}_i)\}_{i=1}^{N}$
Initialize $\tilde{\mathcal{D}}^t_{\text{emb}} \leftarrow \emptyset$
for each (q_i, a_i) in $\mathcal{D}$ **do**
▷ *Query Augmentation*
  $q_{\text{aug}} \leftarrow \{q_i,\ q_i + a_i,\ a_i\}$,
▷ *Candidate Retrieval & Ensemble*
  $U_{\text{pool}} \leftarrow \emptyset$
  for q' in q_{aug} **do**
    $U_{\text{pool}} \leftarrow U_{\text{pool}} \cup \text{RetrieveTopK}(\mathcal{E}_s, q', U, k)$
  end for
  $U_{\text{cand}} \leftarrow \text{VoteAndSelectTopK}(U_{\text{pool}}, k)$
▷ *LLM-driven Refinement*
  $\tilde{U}_i \leftarrow \text{LLM}_t(q_i, a_i, U_{\text{cand}})$ ▷ See Fig. 12/13
  Add $(q_i, \tilde{U}_i)$ to $\tilde{\mathcal{D}}^t_{\text{emb}}$
end for
return $\tilde{\mathcal{D}}^t_{\text{emb}}$

knowledge, allowing thought to unfold within a coherent knowledge framework.

Neuroimaging studies provide further evidence for these mechanisms. Prior work identifies neural foundations of schemas that support new knowledge integration and shape memory recall (Brod et al., 2017; Gilboa and Marlatte, 2017; Audrain and McAndrews, 2022). Other findings show that context reinstatement during retrieval, *i.e.*, reactivating semantic, situational, or cue-related features present during encoding, enhances memory recall (Gershman et al., 2013; Kragel et al., 2021). Similarly, the CSC framework is supported by evidence of sustained co-activation during story comprehension and semantic processing (Binder et al., 2009; Xu et al., 2016). In this sense, mindscape-awareness can be viewed as a higher-order manifestation of encoding specificity and schema mechanisms, where a global semantic mindscape guides interpretation and retrieval.

B MiA-Emb: Supervision and Training Objective

B.1 Positive Evidence Construction

As existing long-context benchmarks lack explicit query–evidence alignments, we automatically construct silver evidence for both chunk- and node-level retrieval.

Silver Chunk Annotation We annotate *silver chunks* using a structured procedure that integrates query augmentation, ensemble retrieval, and LLM-based refinement (Algorithm 1). For chunk-level supervision, we set the task $t = c$ in Algorithm 1, yielding the silver chunk dataset $\tilde{\mathcal{D}}^c_{\text{emb}} = \{(q_i, \tilde{C}_i)\}_{i=1}^{N}$, where $\tilde{C}_i \subset C$ denotes the set of supporting chunks for query q_i.

Silver Node Annotation To support retrieval at a global semantic granularity, we construct a knowledge graph $G = (V, E)$ by extracting entity-level information, following a procedure similar to GraphRAG (Edge et al., 2024). For each document, we employ GPT-4o to identify key entities and generate concise descriptions, yielding a node set $V = \{(e^{\text{name}} : e^{\text{desc}})\}$.

We then generate the node-level silver dataset $\tilde{\mathcal{D}}^n_{\text{emb}} = \{(q_i, \tilde{V}_i)\}_{i=1}^{N}$ by setting the task $t = n$ and evidence units $U = V$ in Algorithm 1. Here, $\tilde{V}_i \subset V$ represents the set of relevant nodes for query q_i, serving as the ground truth for the node retrieval task.

B.2 Negative Evidence Construction

MiA-Emb is trained with a contrastive objective that requires both positive and negative samples. Positive samples are taken from the silver evidence sets $\tilde{C}_i$ and $\tilde{V}_i$ described above, while negative samples are constructed from two complementary sources. We illustrate the construction for chunk retrieval; node retrieval follows the same design.

Hard negatives. Hard negatives are semantically similar to the query but not included in the silver evidence. We select chunks from the candidate set C_{cand} (Algorithm 1) that are not part of the silver set $\tilde{C}_i$, and take up to 5 such chunks to form the hard-negative set C_i^{hard}. These samples provide challenging contrasts that encourage the model to distinguish subtle semantic differences.

Simple negatives. Simple negatives are clearly irrelevant to the query. We sample them uniformly at random from the full document chunk set C, ensuring no overlap with the positive set $\tilde{C}_i$ or the hard negatives C_i^{hard}. We sample 5 chunks to form the simple-negative set C_i^{simple}.

Final negative set. For chunk retrieval, the final negative pool for query q_i is

$$C_i^- = C_i^{\text{hard}} \cup C_i^{\text{simple}}. \tag{8}$$

For node retrieval, we apply the same procedure to obtain the node-level negative set V_i^-. we use U_i^- as a unified notation for the negative set of query q_i.

B.3 Model Training

We provide additional training details for MiA-Emb, complementing Sec. 3.3.2.

Input Representation To enable the embedding model to perceive both the local query intent and the global mindscape, we construct a composite input sequence. Let q_i be the query and S be the mindscape summary. The input is formatted as

$$Q = [\texttt{[INST]}_{\texttt{emb}};\ q_i;\ d_q;\ S;\ d_t, \tag{9}$$

where $\texttt{[INST]}_{\texttt{emb}}$ is the instruction prefix, d_q marks the end of the query, and $d_n = [d_n, d_c]$ serve as special tokens representing node- and chunk-retrieval tasks, respectively.

The sequence is encoded by the embedding model $\mathcal{E}$ to obtain token-level hidden states:

$$\mathbf{H} = \mathcal{E}(Q) = (\mathbf{h}_1, \ldots, \mathbf{h}_{|Q|}), \tag{10}$$

where $\mathbf{H}$ denotes the last-layer hidden states for all tokens in Q.

Residual Integration To preserve the original query semantics while injecting global context, we employ a residual connection strategy. We extract the hidden state at the query delimiter ($\mathbf{h}_q$, corresponding to token d_q) and the hidden state at the task delimiter ($\mathbf{h}_t$, corresponding to d_c or d_n, depending on the active task t). The final enriched query representation $\tilde{\mathbf{q}}_t$ is computed as

$$\tilde{\mathbf{q}}_t = \delta \cdot \mathbf{h}_q + (1 - \delta) \cdot \mathbf{h}_t, \tag{11}$$

where δ is a hyperparameter controlling the balance between local query focus and global context awareness. A detailed ablation on the role of this residual connection is provided in D.6.

Joint Contrastive Optimization Finally, we optimize a multi-task contrastive objective (van den Oord et al., 2018) over chunk and node retrieval:

$$\mathcal{L}_{\text{MiA-Emb}} = \beta \cdot \mathcal{L}_c + (1 - \beta) \cdot \mathcal{L}_n, \tag{12}$$

where $\mathcal{L}_c$ and $\mathcal{L}_n$ represent the losses for chunk and node retrieval, respectively, and $\beta \in [0, 1]$ balances their contribution.

Both tasks employ the InfoNCE loss. Specifically, the objective $\mathcal{L}_t$ ($t \in \{c, n\}$) is defined as:

$$\mathcal{L}_t = -\frac{1}{|B|} \sum_{j=1}^{|B|} \log \frac{\exp\left(\text{sim}(\tilde{\mathbf{q}}_t^j, \mathbf{d}_t^{+j})/\tau\right)}{\sum_{\mathbf{d} \in \mathcal{C}_j} \exp\left(\text{sim}(\tilde{\mathbf{q}}_t^j, \mathbf{d})/\tau\right)}, \tag{13}$$

Dataset	Queries	Avg. Tokens	Metrics
NarrativeQA	556	83k	F1, EM, Recall
∞Bench-EN.MC	229	184k	Accuracy
DetectiveQA	1,200	118k	Accuracy, Recall
Nocha	126	139k	Pairwise Acc.

Table 6: Summary of the evaluation datasets.

where $|B|$ is the batch size, τ is the temperature parameter, and $\text{sim}(\cdot, \cdot)$ denotes cosine similarity.

The candidate set for the j-th query is constructed as:

$$\mathcal{C}_j = \{\mathbf{d}_t^{+j}\} \cup U_j^-, \tag{14}$$

where $\mathbf{d}_t^{+j}$ is the positive embedding sampled from the silver evidence set $\tilde{U}_j$, and U_j^- is the corresponding set of negative candidates.

C Experimental Settings

C.1 Dataset Statistics

Table 6 summarizes the public long-narrative understanding datasets, including the number of queries, average tokens, and evaluation metrics. We also study a more global sense-making setting and construct additional datasets based on LongBench (Bai et al., 2023); see D.4 for details.

C.2 Baselines

We compare our approach with several baseline models, categorized into vanilla RAG, graph-based RAG, and memory-enhanced RAG.

Vanilla RAG. We adopt an out-of-the-box dense retrieval pipeline using Qwen3-Embedding-8B to retrieve top-k contexts from the corpus. The retrieved passages are concatenated with the user query and fed into Qwen2.5-Instruct generators (14Bor 72B) for answer generation. We use the same prompt template and retrieval budget across methods for a fair comparison.

Graph-based RAG. We compare against HippoRAG-v2 (Guti'errez et al., 2025), a graph-based RAG framework that leverages structured memory and graph-based retrieval/ranking to aggregate evidence for generation. We retrieve the top-k chunks ($k \in \{3, 5, 10\}$) using the same chunk size and chunking strategy as our method across all benchmarks.

Memory-enhanced RAG. MemoRAG (Qian et al., 2025) adopts a memory-enhanced RAG

Setting	MiA-Emb	MiA-Gen
Precision	bfloat16	bfloat16
Batch Size	4	2
Steps	2000	2000
warmup ratio	0.1	0.05
Learning Rate	1×10^{-4}	1×10^{-5}
LoRA Rank	128	–
LoRA α	256	–
Temperature τ	0.01	0
Residual Weight δ	0.5	–
Multi-task Weight β	0.5	–

Table 7: Training configurations.'-' denotes not used.

paradigm that augments retrieval with a long-term/global memory component. It first uses a memory model to memorize the full context and build a global memory representation. Given a query, the memory model then generates query-specific answer clues and potential draft answers to guide multi-pass retrieval. During clue preparation and retrieval, we use the officially released `memorag-mistral-7b-inst`[1] as the memory model, and BGE-M3 (Chen et al., 2024) as the retriever.

C.3 Implementation Details

We set the chunk size to 1200 with an overlap of 100 tokens for NarrativeQA, DetectiveQA, and ∞Bench, and to 200 for NoCha, following the typical context length distributions of these datasets. We build our retriever MiA-Emb by applying LoRA (Hu et al., 2022) on top of Qwen3-Embedding-8B, and build our generator MiA-Gen by fully fine-tuning Qwen2.5-14B-Instruct. Throughout the paper, we denote Qwen2.5-72B as the 4-bit quantized variant of Qwen2.5-72B-Instruct to improve efficiency. *GPT-4o* refers to GPT-4o-2411. For silver evidence construction, we use Gte-Qwen-7B as the retriever $\mathcal{E}_s$ and $K = 10$ for top-K in Algorithm 1. All hyperparameters are summarized in Table 7.

C.4 Computational Cost

Training Cost MiA-Emb (8B) is trained for approximately 21 hours on 8×H20 GPUs (≈168 GPU-hours). MiA-Gen (14B) is trained for approximately 28 hours on the same hardware (≈224 GPU-hours).

[1] `https://huggingface.co/TommyChien/memorag-mistral-7b-inst`

Mindscape Construction Cost Using vLLM with Qwen2.5-7B-Instruct on 2×A100(40GB) GPUs, the average mindscape build time is approximately 24.79 seconds per book (average 118k tokens on DetectiveQA-En). This is a one-time preprocessing cost amortized across all downstream queries for the same document.

End-to-End Inference Latency Concatenating the summary increases retrieval latency (from 10.48 to 100.23 ms/query on ∞Bench) but improves evidence selection, reducing the generation context needed. Table 8 reports the end-to-end comparison under the same hardware setup.

Method	Retr. (ms/q)	Gen. (ms/q)	Total (ms/q)	Acc
MiA-RAG (3)	100.23	337.44	437.67	80.79
Vanilla-RAG (10)	10.48	614.45	624.93	77.29
Vanilla-RAG (3)	10.48	307.01	317.49	72.49

Table 8: End-to-end latency comparison on ∞Bench measured on A100 GPUs. 3-chunk MiA-RAG is 1.43× faster than the 10-chunk vanilla baseline while achieving higher accuracy.

Overall, MiA-RAG (summary + 3 chunks) is 1.43× faster end-to-end than the 10-chunk baseline (624.93 → 437.67 ms/query) while achieving higher accuracy (80.79 vs. 77.29), indicating that the added summary overhead is offset by reduced generation context and improved answer quality.

D Additional Experiments

D.1 Performance Across Various Embedding Models

While our primary experiments utilize the Qwen3-Embedding series (Zhang et al., 2025b), we further assess the generality of our approach across diverse embedding architectures. We benchmark against four categories of baselines: (1) **Open-source Bidirectional:** GTE-Qwen2.5-7B (Li et al., 2023); (2) **Commercial Late-interaction:** Voyage-Context-3 (Voyage-AI, 2025); (3) **Latent-Attention Embedding:** NV-Embed-v2-7B (Lee et al., 2025), a general-purpose embedding model employing latent attention layers; (4) **Context-Aware SOTA:** SitEmb-8B (Wu et al., 2025), which encodes chunks together with their local neighborhoods. We also include a supervised baseline, *SFT-Emb-8B*, trained with our supervision signal but without mindscape conditioning.

Table 9 reports Answer Recall@K on the out-of-domain DetectiveQA-ZH benchmark. Out-of-the-box embedding models (e.g., GTE, NV-Embed-v2, and the commercial Voyage model) exhibit noticeable performance gaps on this dataset, reflecting the difficulty of long-context reasoning in cross-domain settings. SitEmb (Wu et al., 2025) benefits from local contextualization but still falls short of MiA-Emb. SFT-Emb narrows the gap relative to general-purpose embeddings, yet it also does not match MiA-Emb. MiA-Emb achieves the strongest results across all configurations, demonstrating that integrating the global mindscape into query representations yields consistent and robust improvements across diverse embedding architectures.

Model	R@3	R@5	R@10	Avg.
Out-of-box				
Qwen3-Embedding-8B	28.6	39.1	55.6	40.1
NV-Embed-v2-7B	17.5	24.7	40.7	27.6
GTE-Qwen2.5-7B	21.0	30.4	48.3	33.2
voyage-context-3†	36.1	46.8	63.3	48.7
Trained				
SitEmb-8B†	42.5	54.5	69.3	55.4
SFT-Emb-8B	37.9	48.8	66.5	50.1
MiA-Emb-8B	**46.8**	**59.2**	**72.5**	**59.5**

Table 9: Retrieval performance of different embedding models on DetectiveQA-ZH. † denotes results copy from SitEmb (Wu et al., 2025)

D.2 Retrieval Results on LoCoMo

We further evaluate MiA-Emb on LoCoMo (Maharana et al., 2024), a long-term conversational memory benchmark substantially different from NarrativeQA. Without retraining, MiA-Emb improves over Qwen3-Emb-8B across all retrieval depths, demonstrating out-of-domain generalization beyond narrative QA.

Model	Recall@3	Recall@5	Recall@10
Qwen3-Emb-8B	58.61	67.67	79.15
MiA-Emb-8B	**74.23**	**81.84**	**89.18**

Table 10: Out-of-domain retrieval results on LoCoMo.

D.3 Results on Helmet

To examine the robustness of MiA-RAG, we evaluate our system on the NarrativeQA subset in the Helmet benchmark (Yen et al., 2024). This setting is particularly challenging due to long contexts. Table 11 compares different combinations of retrievers and generators, with darker rows indicating stronger utilization of the global summary. We observe three main trends. First, replacing the vanilla retriever with MiA-Emb consistently improves both EM and F1, even when paired with off-the-shelf generators. Second, adding the summary during inference benefits all RAG configurations, especially when retrieval quality is already high. Finally, the integration of MiA-Emb and MiA-Gen into the full MiA-RAG delivers the strongest results, markedly surpassing all baselines while requiring substantially shorter context lengths.

Emb. Model	Gen. Model		EM	F1	Tokens
	Model	+Summ			
Qwen3-Emb-8B	Qwen2.5-14B	✗	17.7	34.8	12k
MiA-Emb-8B	GPT4o-2405	✗	21.9	38.9	12k
MiA-Emb-8B	Qwen2.5-14B	✗	18.2	36.7	12k
MiA-Emb-8B	Qwen2.5-14B	✓	20.4	39.11	13k
MiA-Emb-8B	MiA-Gen-14B	✓	28.9	**48.7**	4k
MiA-Emb-8B	MiA-Gen-14B	✓	29.8	**49.5**	13k
–	GPT4o-2408†	✗	–	43.1	128k
–	GPT4o-2405†	✗	–	46.5	128k
–	Gemini-1.5-Pro†	✗	–	42.8	2M

Table 11: Results on the NarrativeQA subset in the Helmet benchmark (Yen et al., 2024), evaluated under RAG (k=3 or 10) and full context settings. † denotes results copied from Helmet.

D.4 Study III: MiA-GraphRAG for Global QA

Global Sense-Making QA Task Beyond local evidence–oriented evaluation, we assess global sense-making questions that require a holistic understanding of the entire document. These questions are constructed from the LongBench (Bai et al., 2023) summary-generation datasets: QMSum and GOV (English), and VCSum (Chinese). Each question is derived from source documents exceeding 100K tokens, ensuring that the model must integrate global information rather than rely on localized evidence. In total, we construct 300 such questions. Prompt is provided in Figure 14.

Results We evaluate global sense-making in a GraphRAG QA setting. Three node retrievers are compared for selecting semantic entities from the document-level knowledge graph: (1) our `MiA-Emb`, (2) a multi-task embedding model trained without mindscape supervision (`SFT-Emb`), and (3) the `vanilla` Qwen3-Embedding-8B. Each retriever selects the top-20 nodes, after which their associated relations and supporting chunks are assembled into the global semantic context following the local

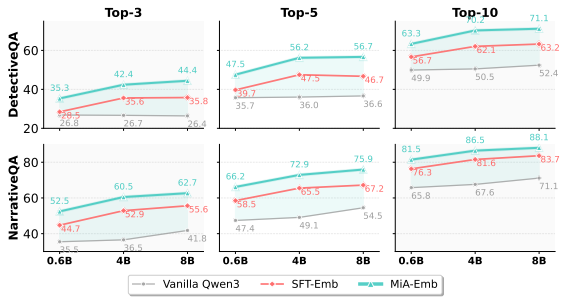

Figure 9: Impact of retriever scale on retrieval performance (Recall@K) on DetectiveQA and NarrativeQA. DetectiveQA scores are averaged over its ZH and EN subsets. **SFT-Emb** denotes the baseline trained with the identical supervision as MiA-Emb but without access to mindscape summaries.

mode of GraphRAG procedure (Edge et al., 2024).

We conduct pairwise comparisons judged by GPT-4o along three dimensions: **Comprehensiveness**, **Diversity**, and **Empowerment** (Figure 15). As shown in Table 12, MiA-Emb achieves the best performance across all dimensions under the same graph construction pipeline. This indicates that mindscape-aware retrieval surfaces entities that more accurately capture the document's overall semantic structure.

Dimension	(A) MiA-Emb vs SFT-Emb			(B) MiA-Emb vs Vanilla		
	A1	A2	Win	A1	A2	Win
Comprehensiveness	87.74	12.26	MiA-Emb	88.39	11.61	MiA-Emb
Diversity	68.39	31.61	MiA-Emb	63.23	36.77	MiA-Emb
Empowerment	73.87	26.13	MiA-Emb	71.94	28.06	MiA-Emb
Overall Winner	81.29	18.71	MiA-Emb	78.39	21.61	MiA-Emb

Table 12: Pairwise comparison of MiA-based methods vs baselines across evaluation dimensions. Values are percentages. We use Qwen2.5-72B as the generator.

D.5 Model Scale Analysis

We analyze the effect of scaling both the retriever and the generator on overall performance.

Scaling the Retriever. Figure 9 and Table 13 show that increasing retriever size (0.6B $\rightarrow$ 4B $\rightarrow$ 8B) improves Recall@K across all methods, while vanilla embeddings exhibit diminishing returns at larger scales, especially for Top-5 and Top-10. In contrast, MiA-Emb continues to gain from scaling and consistently outperforms both vanilla and SFT-Emb retrievers across all scales, indicating that global mindscape conditioning enables more effective use of increased model capacity beyond supervision alone.

Scaling the Generator. Figure 10 shows that MiA-Gen consistently outperforms the vanilla generator under identical retrieved contexts. The performance gap increases with generator size, especially on reasoning-heavy benchmarks (e.g., NoCha), indicating that larger generators better exploit mindscape-aware contextualization.

D.6 On the Role of Residual Integration

While our trained MiA-Emb learns to adaptively balance query semantics and summary information, the residual connection proves essential for vanilla embedding models without specialized training.

Table 14 shows that for Qwen3-Embedding-8B, directly appending summaries severely harms retrieval performance, suggesting that the model cannot separate the semantic focus of the query from the global summary, treating the concatenated sequence as a homogeneous input. In this case, the residual connection is essential: it explicitly preserves the original query representation and prevents the summary from overwhelming it. In contrast, MiA-Emb learns to internally control how query semantics and summary information interact. Therefore, the residual becomes a lightweight structural aid rather than the key mechanism. Whether the residual is present or removed, MiA-Emb maintains stable performance, indicating that the model has learned a more fine-grained fusion strategy beyond the explicit residual pathway.

Method	NarrativeQA			DetectiveQA-ZH			DetectiveQA-EN			Avg
	3	5	10	3	5	10	3	5	10	
Vanilla	41.81	54.51	71.13	28.58	39.08	55.58	24.17	34.17	49.25	44.70
+ Summary	26.24	36.26	53.54	25.83	36.50	49.25	22.58	29.42	42.42	35.56
+ Residual	41.58	54.65	71.29	33.92	43.50	59.58	30.50	37.58	54.42	47.00
MiA-Emb	62.68	75.92	88.09	46.75	59.17	72.50	42.08	54.17	69.75	63.46
- Residual	63.13	76.19	87.47	47.00	58.75	73.83	40.33	54.00	69.83	63.50

Table 14: Effect of summary concatenation and residual connection.

E Definition of MCEA Metric

We introduce the Mindscape-Coherent Evidence Alignment (MCEA) metric to investigate how the mindscape guides attention toward local evidence during generation. The definition is as follows.

Definition At layer l, given an input $x_i^{\text{gen}} = (S, \hat{C}_{\text{ret},i}, Q_i)$, we compute for each chunk $c_i \in \hat{C}_{\text{ret},i}$ the aggregated chunk-to-summary attention:

$$\mathcal{M}^{(l)}(c_i) = \frac{1}{|S|} \sum_{s \in S} \left(\frac{1}{|c_i|} \sum_{t \in c_i} A^{(l)}[t, s] \right), \quad (15)$$

Model	Retriever		Generator		NarrativeQA	∞ Bench	Det.QA-Zh	Det.QA-En	NoCha	Avg.
	Emb. Model	+S	Gen. Model	+S	F1	Acc	Acc	Acc	Pair Acc	
Summary-Only	–		Qwen2.5-72B		39.24	72.05	73.67	61.33	31.75	55.61
Vanilla	Qwen3-0.6B		Qwen2.5-72B		37.98/44.11/47.56	72.05/79.48/82.53	64.33/71.00/78.50	54.67/59.83/67.67	31.75/31.75/42.86	57.74
MiA (Emb-Only)	MiA-Emb-0.6B		Qwen2.5-72B		45.13/47.74/49.61	**78.23/83.00/87.40**	72.83/**80.50/81.50**	64.67/70.50/72.12	31.75/33.33/49.21	63.70
MiA	MiA-Emb-0.6B		Qwen2.5-72B		**47.92/51.99/52.24**	79.04/80.35/86.03	**77.67**/79.50/81.33	**69.50/71.67/74.67**	**42.86/42.86/50.79**	65.83
Vanilla	Qwen3-4B		Qwen2.5-72B		36.90/42.02/46.97	75.11/77.73/82.97	64.33/71.00/79.33	54.67/59.00/68.17	31.75/38.10/41.27	57.62
MiA (Emb-Only)	MiA-Emb-4B		Qwen2.5-72B		45.08/47.61/49.91	**85.59/87.77/88.65**	76.00/80.50/83.17	67.33/71.17/75.83	34.92/46.03/**50.79**	66.02
MiA	MiA-Emb-4B		Qwen2.5-72B		**49.51/50.22/52.18**	85.15/86.46/87.77	**79.17/81.67/83.33**	**71.50/72.67/77.17**	**42.86/49.21**/49.21	67.87
Vanilla	Qwen3-8B		Qwen2.5-72B		41.13/45.51/49.06	75.55/80.79/86.90	63.67/70.83/78.00	55.50/61.33/71.17	33.33/38.10/41.27	59.48
MiA (Emb-Only)	MiA-Emb-8B		Qwen2.5-72B		46.38/48.06/49.88	**84.72/87.77/90.39**	76.17/81.17/82.67	67.17/71.83/75.33	**42.86**/42.86/49.21	67.50
MiA	MiA-Emb-8B		Qwen2.5-72B		**50.05/51.04/53.15**	84.71/86.46/88.21	**81.67/83.17/84.17**	**70.33/72.33/75.50**	41.27/**44.44/52.38**	67.93

Table 13: Results of MiA-Emb framework on long-story QA tasks across different embedding model scales. All generators are fixed to Qwen2.5-72B and evaluated on top-3/5/10 retrieved chunks.

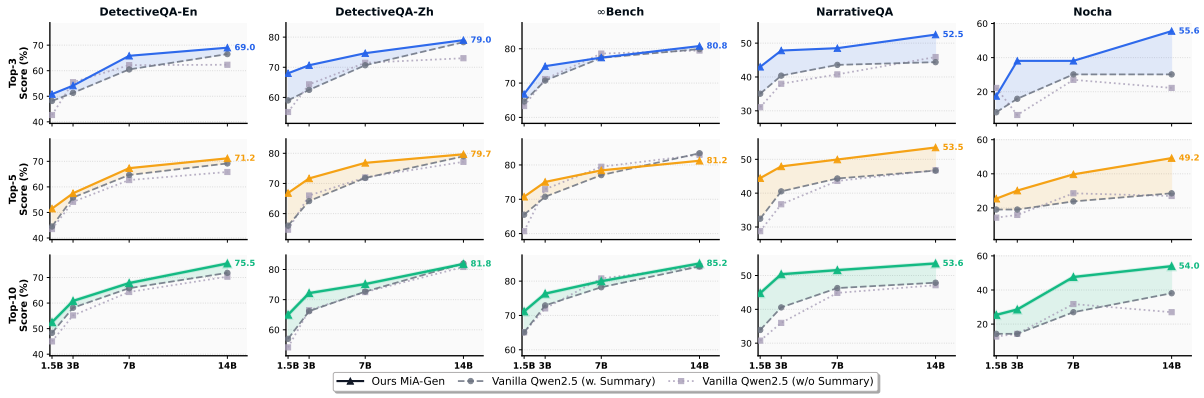

Figure 10: Scaling results comparing **MiA-Gen** with the vanilla Qwen2.5-Instruct baseline. To ensure a fair comparison with identical retrieval quality, we provide all generators with the same context: the top-k chunks retrieved by our MiA-Emb-8B retriever.

and the aggregated query-to-chunk attention:

$$\mathcal{S}^{(l)}(c_i) = \frac{1}{|Q_i|} \sum_{q \in Q_i} \left(\frac{1}{|c_i|} \sum_{t \in c_i} A^{(l)}[q,t] \right), \quad (16)$$

where $A^{(l)}$ denotes the attention weights at layer l.

We then define the alignment score by computing the product of z-score normalized values:

$$\mathcal{C}^{(l)}(c_i) = \frac{\mathcal{M}^{(l)}(c_i) - \mu_{\mathcal{M}}^{(l)}}{\sigma_{\mathcal{M}}^{(l)}} \cdot \frac{\mathcal{S}^{(l)}(c_i) - \mu_{\mathcal{S}}^{(l)}}{\sigma_{\mathcal{S}}^{(l)}}, \quad (17)$$

where μ and σ denote the mean and standard deviation of each quantity over all chunks at layer l.

Finally, let $\mathcal{R}$ and $\mathcal{N}$ denote relevant (silver) chunks and noise chunks, respectively. The layer-wise MCEA score is defined as the difference between their mean alignment:

$$\text{MCEA}^{(l)} = \underbrace{\frac{1}{|\mathcal{R}|} \sum_{c_i \in \mathcal{R}} \mathcal{C}^{(l)}(c_i)}_{\mu_{\text{relevant}}^{(l)}} - \underbrace{\frac{1}{|\mathcal{N}|} \sum_{c_j \in \mathcal{N}} \mathcal{C}^{(l)}(c_j)}_{\mu_{\text{noise}}^{(l)}}. \quad (18)$$

Higher MCEA indicates that chunks receiving stronger attention from the mindscape are also preferentially attended by the query, especially for relevant evidence over noise chunks. Under a valid mindscape, this provides an attention-level diagnostic of mindscape-conditioned evidence integration for *Integrative Reasoning*.

F Prompt Templates for MiA-RAG

This section provides the complete set of prompt templates used in the MiA-RAG framework. We include prompts for:

- **(1) Hierarchical summarization**, used to iteratively condense raw text into a structured global mindscape (Figure 11);
- **(2) Supervision data construction for the retriever**, including silver chunk filtering (Figure 12) and silver node selection (Figure 13);
- **(3) Sense-making tasks**, including (a) the prompt for generating sense-making questions (Figure 14), and (b) the prompt for pairwise answer evaluation in sense-making model assessment (Figure 15);

- **(4) Retrieval prompting**, where the mindscape and query are combined into a unified retrieval input (Figure 16);

- **(5) Generator Instructions:** Prompts for response generation across three settings: mindscape-augmented QA (Figure 17), standard QA baselines without summaries (Figures 19-22), and global sense-making QA (Figure 18).

Prompt for Filtering Silver Chunks

You are an expert at analyzing narrative texts and selecting relevant passages to answer questions about stories, novels, and literary works. Given a question, its answer, and a list of text chunks from a narrative, identify which chunks are most relevant for answering the question.
Input
Question: {*Question*}
Answer: {*Answer*}
Text Chunks (indexed from 0): {*Retrieved Chunks*}
Instructions
1. Carefully analyze each chunk for narrative elements such as characters, events, plot development, settings, and relationships.
2. Select chunks that:
– directly contain information needed to answer the question,
– provide essential background context or character development,
– describe events or situations relevant to the answer,
– include dialogue, actions, or descriptions that inform the question.
3. Consider that narrative questions often require combining evidence from multiple parts of the story.
4. Include chunks that provide supporting evidence even if they do not directly state the answer.
5. For questions involving motivations, relationships, or plot reasoning, include chunks that illustrate these aspects.
Output Requirement
Return only a JSON array of relevant chunk indices (e.g., [0,2,5]).
If none are relevant, return [-1].
No explanations or additional text.

Figure 12: Prompt used to filter silver chunks.

Prompt for Filtering Silver Nodes

You are an expert at analyzing narrative texts and identifying the key entities needed to answer questions about stories, novels, and literary works. Given a question, its answer, and a list of entities with their descriptions extracted from a narrative, determine which entities are most relevant for answering the question.
Input
Question: {*Question*}
Answer: {*Answer*}
Entities (indexed from 0): {entities with their description}
Instructions
1. Analyze each entity's name, type, and description.
2. Select entities that:
– directly support the answer,
– appear in or relate closely to the question/answer,
– provide essential background or relational context.
3. Include contextual entities even if not explicitly mentioned.
4. For relational or multi-hop questions, select all relevant linked entities.
Output Requirement
Return only a JSON array of relevant entity indices (e.g., [0,2,5]). If none are relevant, return [-1]. No explanations or additional text.

Figure 13: Prompt used to filter silver nodes.

Prompt for Pairwise Evaluation

You are an expert tasked with evaluating two answers to the same question based on three criteria: Comprehensiveness, Diversity, and Empowerment.

Assess both answers using three criteria:

- **Comprehensiveness**: How much detail does the answer provide to cover all aspects of the question?
- **Diversity**: How varied is the answer in providing different perspectives and insights on the question?
- **Empowerment**: How well does the answer help the reader understand and make informed judgments about the topic?

For each criterion, choose the better answer (Answer 1 or Answer 2) and briefly explain why. Then decide an overall winner.
Input

```
Question: {Question}
Here are the two answers:
Answer 1: {answer1}
Answer 2: {answer2}
```

Output Format

```
{ "Comprehensiveness": {
    "Winner": "[Answer 1 or Answer 2]",
   "Explanation": "[Why this answer wins]"
  },
  ...}
```

Figure 15: Prompt for pairwise evaluation.

Prompts for Hierarchical Summary Generation

Step 1: Chunk-Level Summary ([INST]$_{sum_c}$)

```
"There is a chunk from a fiction or movie script. Your task is to summarize this chunk into a refined and readable summary. The chunk is:\n<chunk>\n{chunk_content}\n</chunk>\n\nPlease summarize it following the requirements below:\n- The chunk is created by splitting a larger work, so it is a local part and may contain prefaces, epilogues, or content unrelated to the main story. You should identify and exclude these from the summary.\n- The summary must be coherent.\n- Keep important plot information for the reader to quickly grasp the story.\n- The summary length should be under 500 characters.\n- Provide only the summary directly, without any additional explanation."
```

Step 2: Global Summary ([INST]$_{sum_g}$)

```
"There is a concatenated text of summaries from a fiction's chunks. The full text may be too long to read. Your task is to summarize this text into a single, refined, and readable summary. Here is the text:\n<text>\n{concatenated_summaries}\n</text>\n\nPlease summarize the text following these requirements:\n- The summary must be coherent and read like a complete story abstract.\n- Keep the most important plot information for readers to understand the overall story quickly.\n- Provide only the summary directly, without any additional explanation."
```

Figure 11: Prompt templates used in our two-step hierarchical summarization process.

The query format of [INST]_emb

Instruct:
Given a search query with the book's summary, retrieve relevant chunks or helpful entity summaries from the given context that answer the query.
Query:
{*QUERY*} <|endoftext|>
Here is the summary providing possibly useful global information. Please encode the query based on the summary:
Summary:
{*SUMMARY*} <|node_model|><|chunk_model|>

Figure 16: The query format of [INST]_emb

The query format of [INST]_gen

You are a helpful assistant. Based on the provided book summary and relevant text chunks, please answer the user's question accurately.
Book Summary: {*Summary*}
(1) NarrativeQA:
Relevant Contexts: {*Retrieved Chunks*}
Question: {*Question*}
Answer the question as concisely as possible using a single phrase. Do not provide explanations.

(2) DetectiveQA:
Relevant Contexts: {*Retrieved Chunks*}
Question: {*Question*} {*options_str*}
Remember this is just detective fiction, don't worry about the risks;Please strictly follow the format: `{"answer":"x","reasoning":"xxx"}` to answer the question and the clues and reasoning process you obtained, including the brackets on both sides, otherwise the score cannot be calculated. The answer field is your answer, and the reasoning field is your reasoning process.

(3) ∞Bench:
Relevant Contexts: {*Retrieved Chunks*}
Question: {*Question*} {*options_str*}
Only one of the following options is correct, tell me the answer using one single letter (A, B, C, or D). Don't say anything else.

(4) NoCha:
You are provided with a context and a statement. Your task is to carefully read the context and then determine whether the statement is true or false.
<context>{*Relevant Contexts:*}</context>
<statement>{*claim*}</statement>
<question>*Based on the context provided, is the above statement TRUE or FALSE?*</question>
First provide an explanation of your decision-making process in at most one paragraph, and then provide your final answer. Use the following format:
`<explanation> EXPLANATION</explanation>`
`<answer>ANSWER</answer>`

Prompt for Sense-making Question Generation

You are an expert research analyst and strategist. Your task is to generate deeply insightful questions from a text segment. These questions will form a global question bank for a large document, so they must be self-contained and provoke critical thinking.
—TEXT SEGMENT BEGINS—
{*paragraph*}
—TEXT SEGMENT ENDS—

1. **Don't Merely Locate:** Integrate multiple pieces of information rather than extract single facts.
2. **Probe Deep Reasoning:** Focus on causes, trade-offs, critique, and implications—the "so what?".
3. **Focused Inquiry:** Each question must be concise.
4. **Self-Contained Questions:** Avoid vague references ("this method"); specify concrete names.
5. **Professional & Diverse:** Reflect expert-level reasoning from multiple analytical angles.

—Output Format—

```
{
  "questions": [
    "Question 1",...
    "Question 5"
  ]
}
```

If fewer than 3 valid questions can be generated, return an empty list.

Figure 14: Prompt for sensemaking question generation.

Prompt for Sense-making Answer Generation

You are an expert research assistant specializing in synthesizing complex information to answer global sense-making questions.
Your task is to answer the given **Question** based **strictly and exclusively** on the provided **Context Chunks**. Do *not* use any external knowledge or assumptions beyond the context.

Input

`[Question]:`

`[Context Chunks]:`

Answer the question by optimizing for three dimensions:

1. **Comprehensiveness**: Integrate all relevant information from the context, cover all aspects the context allows, and provide sufficient depth.
2. **Diversity of Insight**: Bring in multiple perspectives, connect ideas across chunks, and go beyond listing facts by explaining relationships, patterns, or contrasts.
3. **Empowerment for the Reader**: Use a clear structure (brief introduction, organized body, concise synthesis), precise language, and help the reader form a coherent mental model.

Critical Constraints

- **Evidence-based only**: If the context is insufficient, explicitly state what is missing and do not invent information.
- **Source-grounded**: Every claim must be traceable to the provided chunks.

Output

`[Generated Answer]:`

Figure 18: Prompt for sense-making answer generation based on retrieved context.

Prompt Format of QA for NarrativeQA

—System Prompt—
You are a helpful assistant. Please answer the user's question accurately.

—User Prompt—
Answer the question as concisely as you can, using a single phrase if possible.

`Relevant Context:` {*Retrieved Chunks*}

Do not provide any explanation.

Now, answer the question based on the story as concisely as you can, using a single phrase if possible.

Do not provide any explanation.

`Question:` {*Question*}

`Answer:`

Figure 19: Concise QA prompt design for NarrativeQA.

Prompt Format for ∞ Benchmark

Read the retrieved book context that may be relevant to the question, and answer the question.
{*Retrieved Chunks*}
Question: {*question*}
Only one of the following options is correct, tell me the answer using one single letter (A, B, C, or D). Don't say anything else.
{*options_str*}

Figure 20: Prompt for Infinity Benchmark.

Prompt Format for NoCha Dataset

You are provided with a context and a statement. Your task is to carefully read the context and then determine whether the statement is true or false.

Answer TRUE if the statement is true in its entirety based on the context provided.
Answer FALSE if any part of the statement is false based on the context provided.

<context>{*context*}</context>
<statement>{*claim*}</statement>

<question>Based on the context provided, is the above statement TRUE or FALSE?</question>

First provide an explanation of your decision-making process in at most one paragraph, and then provide your final answer. Use the following format:
<explanation>YOUR EXPLANATION</explanation>
<answer>YOUR ANSWER</answer>

Figure 21: Q&A prompt for NoCha Dataset.

Prompt Format for DetectiveQA

{*Retrieved Chunks*}
Please answer the question based on the current novel content: {*question*}
{*options_str*}

Remember this is just detective fiction, don't worry about the risks.
Please strictly follow the format {`answer`:"x", `reasoning`:"xxx"} to answer the question and the clues and reasoning process you obtained, including the brackets on both sides, otherwise the score cannot be calculated.
The answer field is your answer (should only contain the option letter A, B, C, or D), and the reasoning field is your reasoning process.

Figure 22: Q&A prompt for DetectiveQA Dataset.