ComoRAG: A Cognitive-Inspired Memory-Organized RAG for Stateful Long Narrative Reasoning

Juyuan Wang[1*], Rongchen Zhao[1*], Wei Wei[2], Yufeng Wang[1], Mo Yu[4], Jie Zhou[4], Jin Xu[1,3], Liyan Xu[4†]

[1]School of Future Technology, South China University of Technology
[2]Independent Researcher
[3]Pazhou Lab, Guangzhou
[4]WeChat AI, Tencent

Abstract

Narrative comprehension on long stories and novels has been a challenging domain attributed to their intricate plotlines and entangled, often evolving relations among characters and entities. Given the LLM's diminished reasoning over extended context and its high computational cost, retrieval-based approaches remain a pivotal role in practice. However, traditional RAG methods could fall short due to their stateless, single-step retrieval process, which often overlooks the dynamic nature of capturing interconnected relations within long-range context. In this work, we propose ComoRAG, holding the principle that narrative reasoning is not a one-shot process, but a dynamic, evolving interplay between new evidence acquisition and past knowledge consolidation, analogous to human cognition on reasoning with memory-related signals in the brain. Specifically, when encountering a reasoning impasse, ComoRAG undergoes iterative reasoning cycles while interacting with a dynamic memory workspace. In each cycle, it generates probing queries to devise new exploratory paths, then integrates the retrieved evidence of new aspects into a global memory pool, thereby supporting the emergence of a coherent context for the query resolution. Across four challenging long-context narrative benchmarks (200K+ tokens), ComoRAG outperforms strong RAG baselines with consistent relative gains up to 11% compared to the strongest baseline. Further analysis reveals that ComoRAG is particularly advantageous for complex queries requiring global comprehension, offering a principled, cognitively motivated paradigm for retrieval-based stateful reasoning.

Code — https://github.com/EternityJune25/ComoRAG
Extended version — https://arxiv.org/abs/2508.10419

1 Introduction

The core challenge of long narrative comprehension lies not merely in connecting discrete pieces of evidence, a task more naturally defined as multi-hop Question Answering (QA), but in performing a **dynamic cognitive synthesis** to grasp necessary background and content progression (Xu et al. 2024a). Unlike multi-hop QA (Yang et al. 2018), which seeks a static path through fixed facts, narrative comprehension requires emulating a human reader: continuously building and revising a **global mental model** of the plot, characters, and their evolving motivations (Johnson-Laird 1983). The complexity of this process is well exemplified by a classic question *"Why did Snape kill Dumbledore?"* from the Harry Potter series. Answering this requires weaving a complete web of evidence from disparate clues spanning multiple books—Dumbledore's terminal illness, the Unbreakable Vow, and Snape's deeply concealed loyalty. The true significance of these clues is only fully reconciled in hindsight. This capability is what we term **stateful reasoning**: it demands more than linking static evidence; it requires maintaining a *dynamic memory* of the narrative, one that is constantly updated as new revelations emerge. Long-context LLMs have demonstrated promising performance on benchmarks such as the "Needle in a Haystack" (Eisenschlos, Yogatama, and Al-Rfou 2023). However, their capacity to process long narratives (200k+ tokens) remains limited by finite context windows. Furthermore, as the input length increases, these models are prone to the "lost in the middle" problem (Liu et al. 2024), which raises perplexity and impairs generation quality. This limitation is particularly pronounced in narrative tasks which require stateful reasoning. As a result, retrieval-augmented generation (RAG) (Lewis et al. 2020) has emerged as an important direction for tackling long context comprehension with LLMs, leveraging text embeddings or more advanced retrieval paradigms such as embeddings situated on global context (Wu et al. 2025).

However, existing RAG methods still struggle to effectively address this challenge. Advanced single-step retrieval remains limited by its static index. This includes methods such as RAPTOR (Sarthi et al. 2024), which clusters and summarizes text chunks to retrieve at different levels of details; HippoRAGv2 (Gutiérrez et al. 2025) and GraphRAG (Edge et al. 2025), which build knowledge graphs to achieve multi-hop reasoning in a single retrieval step. Nonetheless, one-shot static retrieval inevitably leads to shallow comprehension. For example, the evidence about Snape in Fig. 1(a)

*These authors contributed equally.
†Project lead. Correspondence to: <liyanlxu@tencent.com>

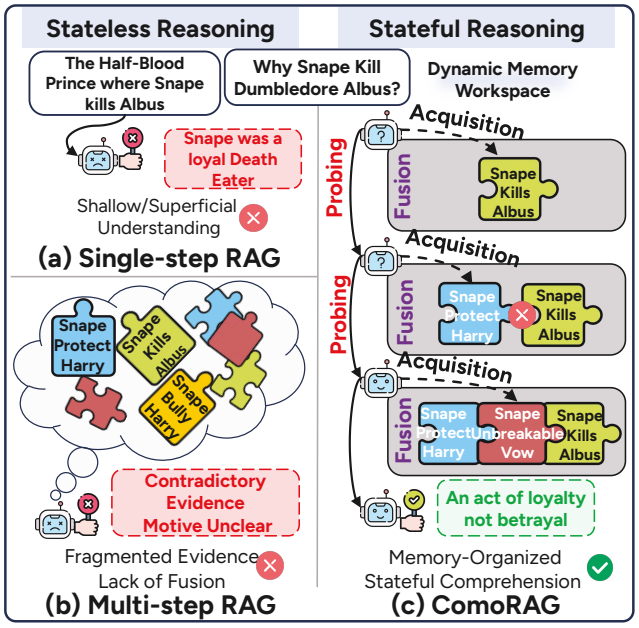

Figure 1: Comparison of RAG reasoning paradigms.

can mislead the model into making a false inference.

As a remedy, multi-step retrieval methods offer a more promising direction, such as IRCoT (Trivedi et al. 2023), which interleaves the retrieval process with Chain-of-Thought reasoning (Wei et al. 2022); Self-RAG (Asai et al. 2024), which trains a model to adaptively retrieve and reflect on evidence; and MemoRAG (Qian et al. 2025), which uses a dual-system architecture to generate clues from compressed global context. These methods all target to obtain richer context through iterative retrieval. However, their retrieval steps are typically independent, which lack coherent reasoning throughout explicit narrative progression, featuring fragmented evidence with a stateless comprehension. As illustrated in Figure 1(b), due to a lack of dynamic memory, multi-step retrieval fails to integrate contradictory evidence such as *"Snape protects/bullies Harry"* and cannot understand the evolution of his actions, ultimately unable to yield the correct answer.

In this work, we seek inspiration from the function of Prefrontal Cortex (PFC) in human brains, which employs a sophisticated reasoning process called **Metacognitive Regulation** (Fernandez-Duque, Baird, and Posner 2000). This process is not a single action but a dynamic interplay between **new evidence acquisition**, driven by goal-directed **memory probes** (Dobbins and Han 2006; Miller and Constantinidis 2024), and subsequent **knowledge consolidation**. During consolidation, new findings are integrated with past information to construct an evolving, coherent narrative. This iterative cycle allows the PFC to continuously assess its understanding and revise its strategy, providing a direct cognitive blueprint for our framework's stateful reasoning approach.

We introduce ComoRAG, a cognitive-inspired, memory-organized RAG framework, imitating the human Prefrontal Cortex (PFC) for achieving stateful reasoning. At its core is a dynamic cognitive loop operating on a memory workspace, which actively probes and integrates new evidence to build a coherent narrative comprehension.

This process, as illustrated in Figure 1(c), is a closed loop of evolving reasoning states. Faced with a complex query like *"Why did Snape kill Dumbledore?"*, the system's memory state evolves from an initial "causally incomplete event" (*Snape kills Albus*), to an "apparent contradiction" upon finding contradictory information (*Snape protects Harry*), and ultimately to a logically consistent **coherent context** through deeper exploration and evidence fusion. Only in this final, complete cognitive state can ComoRAG perform the correct stateful reasoning, deriving the true insight that it was *"an act of loyalty, not betrayal"*.

This cognitively-inspired design yields substantial improvements across four challenging long-context narrative benchmarks. ComoRAG is shown to consistently outperform all categories of strong baselines across each dataset. Our analysis reveals several key findings. First, these gains stem directly from the cognitive loop, which transforms a static knowledge base into a dynamic reasoning engine; for instance, accuracy on EN.MC jumps from a static-retrieval baseline of 64.6% to 72.9%, with performance efficiently converging in around 2-3 cycles. Second, our framework excels on *narrative queries* that require global understanding of plot progression, achieving up to a **19%** relative F1 improvement on these challenging question types where others falter. Finally, our framework demonstrates remarkable modularity and generalizability. Its core loop can be flexibly integrated to existing RAG methods such as RAPTOR, which directly yields a 21% relative accuracy gain). Also, switching to a stronger model as the backbone LLM agents can upgrade reasoning in the entire cognitive loop, attaining accuracy from 72.93% to 78.17%. These results collectively validate that ComoRAG provides a principled, cognitively-inspired new paradigm for retrieval-based long narrative comprehension towards stateful reasoning.

2 Methodology

We introduce ComoRAG, an autonomous cognitive architecture designed to formalize and implement the process of **Metacognitive Regulation** outlined in the Introduction. The architecture's design is directly inspired by the functional mechanisms of the Prefrontal Cortex (PFC) and is founded on three conceptual pillars: (1) a **Hierarchical Knowledge Source** for deep contextual understanding; (2) a **Dynamic Memory Workspace** for tracking and integrating the multi-turn reasoning; and (3) a **Metacognitive Control Loop** that drives the entire resolving procedure.

2.1 Problem Formulation: Towards Principled Narrative Reasoning

Our objective is to design a framework for stateful reasoning in RAG scenarios. Especially, it aims to resolve those queries that require global context comprehension in the first place, commonly seen in narratives, where conventional RAG may fail to recognize relevant context based on the surface form of queries. Formally, denote the initial query as

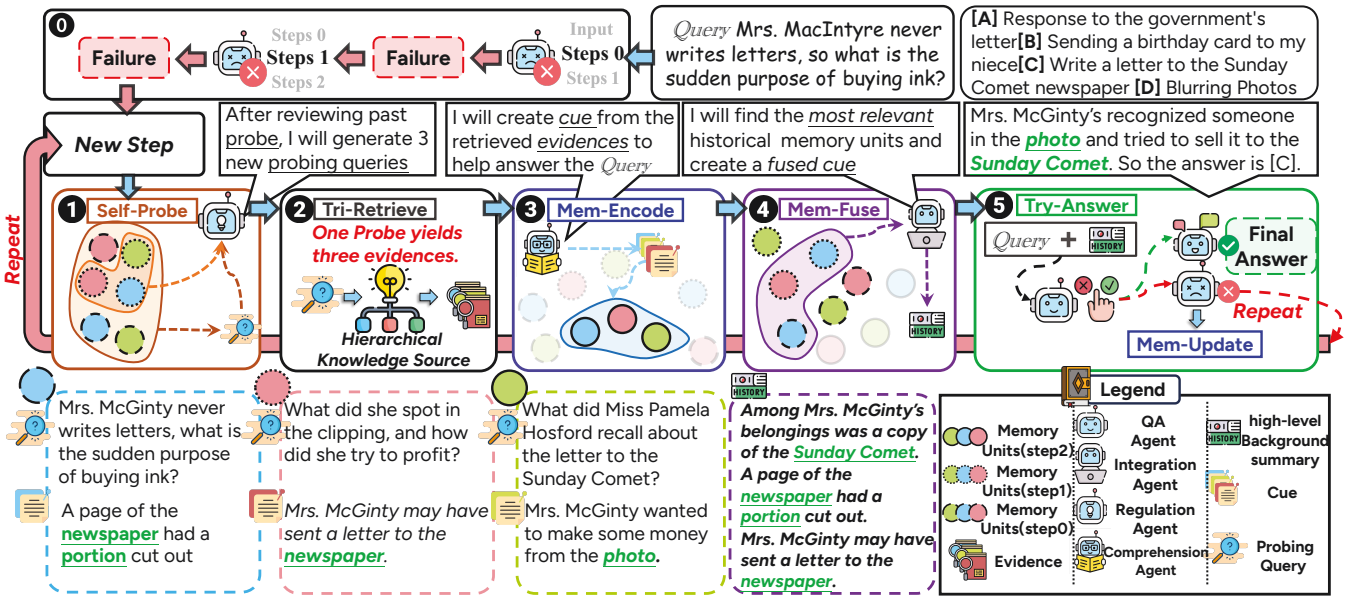

Figure 2: **An illustration of ComoRAG.** Triggered by a reasoning impasse (Failure), the Metacognitive Regulation loop consists of five core operations described in Section 2.3: 1) **Self-Probe** to devise new exploratory probing queries based on past memory units; 2) **Tri-Retrieve** to retrieve evidence from three knowledge sources; 3) **Mem-Encode** to form new memory units on how the latest evidence of new aspects could complement the final query resolution; 4) **Mem-Fuse** to generate cues integrating new and past memory units; 5) **Try-Answer** to perform query answering using new memory information produced in this cycle.

q_{init}, and a knowledge source $\mathcal{X}$ derived upon the original context, our framework F leverages a series of adaptive operations to yield the final answer, A_{final}, through discrete time steps $t = 1, \ldots, T$ with underlying memory control.

At the beginning of each step t, F determines its focus of reasoning—a set of new probing queries $\mathcal{P}^{(t)}$, representing new information to seek that may logically deepen the query comprehension and ultimately complement the answer resolution. With newly retrieved information by $\mathcal{P}^{(t)}$ at each step, the framework utilizes the global memory pool maintained till the prior step $\mathcal{M}_{pool}^{(t-1)}$, and produces either the final answer, or a Failure Signal, indicating a reasoning impasse—and updates the memory pool to $\mathcal{M}_{pool}^{(t)}$, accomplishing a cognitive cycle that synergizes between the knowledge source, memory space and retrieval operations.

2.2 The Hierarchical Knowledge Source

To overcome the limitations of a monolithic representation of the given context, our framework first builds a hierarchical knowledge index $\mathcal{X}$ for retrieval that models the raw text from three complementary cognitive dimensions, analogous to how the PFC integrates different memory types from various brain regions, particularly supporting cross-layer reasoning from raw evidence to abstract relationships.

Veridical Layer: Grounding in Factual Evidence. To ensure all reasoning is traceable to source evidence, a veridical layer $\mathcal{X}^{ver}$ is firstly established, constituted by raw text chunks directly, analogous to the precise recall of factual details in human memory. For more accurate retrieval on text chunks, we instruct a LLM to generate knowledge triples (*subject-predicate-object*) for each text chunk. These triples participate in each retrieval, and strengthen the matching between an incoming query and the corresponding text chunk, which is proven effective by HippoRAG (Jimenez Gutierrez et al. 2024).

Semantic Layer: Abstracting Thematic Structure. To capture thematic and conceptual connections that transcend across long-range contextual dependencies, a semantic layer $\mathcal{X}^{sem}$ is built, inspired by the prior work RAPTOR that employs a GMM-driven clustering algorithm to recursively summarize semantically similar text chunks into a hierarchical summary tree. We reckon such semantic abstraction is necessary for deeper comprehension and follow the same formulism. These summary nodes enable the framework to retrieve conceptual information beyond the surface level.

Episodic Layer: Reconstructing Narrative Flow. The previous two layers equip views of both factual details and high-level concepts. However, they lack temporal development or plot progression that can be especially crucial for narratives. To enable such view with long-range causal chains, we introduce the episodic layer, $\mathcal{X}^{epi}$, which aims to reconstruct the plotline and story arc by capturing the sequential narrative development. The process features a sliding window summarization across text chunks; each resulting node is then a summary that aggregates the narrative development of continuous or causally related events according to the timeline. Optionally, the sliding window process can be applied recursively to form higher-level views of content progression, extracting different levels of narrative flow

as part of the knowledge source.

2.3 The Architecture of Metacognitive Regulation

The core of ComoRAG is a control loop that fully realizes the concept of metacognitive regulation. It is composed of a **Regulatory Process** for reflection and planning at each step, and a **Metacognitive Process** for executing reasoning and memory management with the **Memory Workspace**.

Dynamic Memory Workspace. The memory workspace contains memory units that serve as the bridge for a cohesive multi-step exploration and reasoning by metacognitive regulation. Each memory unit m functionally **concludes one retrieval operation**, denoted as a tuple of three elements: $m = (p, \mathcal{E}_p^{type}, \mathcal{C}_p^{type})$, where p is the *probing query* that triggers this retrieval; $\mathcal{E}_p^{type}$ is the homogeneous set of evidence retrieved from a single knowledge layer ($type \in \{ver, sem, epi\}$); and $\mathcal{C}_p^{type}$ is a *synthesized cue* that reflects how these retrieved evidence by the probe p could complement the comprehension and resolution of the original query q_{init}. Concretely, $\mathcal{C}_p^{type}$ is generated by a LLM in the role of Comprehension Agent, π_{cue}, denoted as $\mathcal{C}_p^{type} = \pi_{cue}(q_{init}, p, \mathcal{E}_p^{type})$.

The formation of a memory unit $(p, \mathcal{E}_p^{type}, \mathcal{C}_p^{type})$ by each retrieval is defined as a **Mem-Encode** operation. The memory workspace/pool will be utilized and updated throughout the reasoning cycle described below.

The Regulatory Process. The regulatory process is invoked at the beginning of a reasoning cycle/step t if the preceding cycle $t-1$ is concluded in failure. The core operation, **Self-Probe**, plans new probing queries of which retrieved information may contribute to the final answer, thereby devising new exploratory paths to break the impasse. It is orchestrated by a **Regulation Agent**, π_{probe}, whose decisions are informed by the reflection on the prior failure, exploring for more necessary background or relevant information towards a full context comprehension to resolve the original query. **Self-Probe** takes three inputs: (1) the ultimate goal q_{init}; (2) the complete exploration probing history $\mathcal{P}_{hist}^{(t-1)}$ up to the end of the last step; and (3) the immediate knowledge gaps that caused the failure, concretized by all *synthesized cues* of memory units generated in the prior step, denoted as $\{\mathcal{C}\}^{(t-1)}$. Its output $\mathcal{P}^{(t)}$ is a new, strategic set of retrieving probes for the current cycle t:

$$\mathcal{P}^{(t)} = \pi_{probe}\big(q_{init},\ \mathcal{P}_{hist}^{(t-1)},\ \{\mathcal{C}\}^{(t-1)}\big) \tag{1}$$

The Metacognitive Process. The metacognitive process takes the new probes for this cycle $\mathcal{P}^{(t)}$, and performs reasoning towards resolving the original query while keeping track of the progress with the memory space. It comprises a series of operations, described in details as follows.

Tri-Retrieve: for each probing query $p \in \mathcal{P}^{(t)}$, a retrieval is conducted on each knowledge layer $\mathcal{X}^{type}$ where $type \in \{ver, sem, epi\}$, such that evidence of high embedding similarity to p per layer is retrieved in a standard Dense Passage Retrieval paradigm, with each evidence being either the raw text chunk, a semantically clustered summary, or a narrative flow summary.

Mem-Encode: for each p and $type$, the retrieved evidence is immediately processed by the aforementioned **Mem-Encode**, to generate a new memory unit that keeps track of how this specific probing could complement to the final answer. The number of all generated memory units at this step can be denoted as $|\mathcal{M}_{encode}^{(t)}| = 3 \times |\mathcal{P}^{(t)}|$.

Mem-Fuse: new memory units in the above step $\mathcal{M}_{encode}^{(t)}$ mainly emphasize aspects probed in the current cycle. To fully utilize the past experience and historical knowledge, the framework further identifies relevant *synthesized cues* from past units in the existing memory pool $\mathcal{M}_{pool}^{t-1}$, then generates a new synthesized cue for fusing past relevant evidence. Let $\mathcal{M}_{pool}^{t-1} \circ q_{init}$ represent past memory units whose cues are of high embedding similarity with q_{init}, and denote a LLM as **Integration Agent** π_{fuse} that synthesizes these relevant past evidence into a high-level background summary, the new cue fusing past memory $\mathcal{C}_{fuse}^{(t)}$ is then:

$$\mathcal{C}_{fuse}^{(t)} = \pi_{fuse}\big(q_{init},\ \mathcal{M}_{pool}^{t-1} \circ q_{init}\big) \tag{2}$$

Try-Answer: with the new probing evidence in $\mathcal{M}_{encode}^{(t)}$ and the past-fusing cue $\mathcal{C}_{fuse}^{(t)}$, a **QA Agent**, π_{QA}, is applied to these contexts to produce the cycle's final output $O^{(t)}$:

$$O^{(t)} = \pi_{QA}\big(q_{init},\ \mathcal{M}_{encode}^{(t)},\ \mathcal{C}_{fuse}^{(t)}\big) \tag{3}$$

Specifically, a LLM is instructed to take these latest evidence and the past background as the context, and determine whether the original query can be resolved. It either yields the **final answer** and terminates the entire reasoning loop, or signals **Failure** and continues to the next step.

Mem-Update: this last step in a cycle simply incorporates the newly generated memory units into the global pool, with their embedding encoded, for future retrieval and reasoning:

$$\mathcal{M}_{pool}^{(t)} \leftarrow \mathcal{M}_{pool}^{(t-1)} \cup \mathcal{M}_{encode}^{(t)} \tag{4}$$

ComoRAG With the above six steps from **Tri-Retrieve** to **Mem-Update**, one cycle of the cognitive loop is realized. For the initial step as in $t = 0$, ComoRAG starts with one round of **Tri-Retrieve** followed by **Try-Answer**. If Failure is signaled, it initiates the Metacognitive loop of stateful reasoning on exploratory paths, characterized by the interlocking operations with the memory workspace, which enables to tackle complex narrative comprehension.

In essence, our framework grasps on the principle that for long context comprehension, especially in narratives where the entire context is cohesively interconnected through the underlying plot progression (Xu et al. 2024a), the query resolution is not a linear pipeline; rather, it is a dynamic, evolving interplay between **new evidence acquisition** and **past knowledge consolidation**, analogous to the human cognitive process. The overall process and detailed prompts used by LLM agents are depicted in our extended paper version.

3 Experimental Settings

Datasets Our experiments cover four long-context narrative understanding datasets for comprehensive evaluation,

Category	Method	NarrativeQA		EN.QA		EN.MC	DetectiveQA	QA Avg.		MC Avg.
		F1	EM	F1	EM	ACC	ACC	F1	EM	ACC
LLM	GPT-4o-mini	27.29	7.00	29.83	12.82	30.57	30.68	28.56	9.91	30.63
Naive RAG	BGE-M3(0.3B)	23.16	15.10	23.71	16.24	59.82	54.54	23.44	15.67	57.18
	NV-Embed-v2 (7B)	27.18	<u>17.80</u>	<u>34.34</u>	<u>24.57</u>	61.13	62.50	30.76	<u>21.19</u>	61.82
	Qwen3-Embed-8B	24.19	15.60	25.79	17.95	<u>65.50</u>	61.36	24.99	16.78	63.43
Enhanced RAG	RAPTOR	27.84	<u>17.80</u>	26.33	19.65	57.21	57.95	27.09	18.73	57.58
	HippoRAGv2	23.12	15.20	24.45	17.09	60.26	56.81	23.79	16.15	58.54
Multi-step RAG	Self-RAG	19.60	6.40	12.84	4.27	59.83	52.27	16.22	5.34	56.05
	MemoRAG	23.29	15.20	19.40	11.64	55.89	51.13	21.35	13.42	53.51
	RAPTOR+IRCoT	<u>31.35</u>	16.00	32.09	19.36	63.76	<u>64.77</u>	<u>31.72</u>	17.68	<u>64.27</u>
	HippoRAGv2+IRCoT	28.98	13.00	29.27	18.24	64.19	62.50	29.13	15.62	63.35
	ComoRAG (Ours)	**31.43**	**18.60**	**34.52**	**25.07**	**72.93**	**68.18**	**32.98**	**21.84**	**70.56**

Table 1: Evaluation results on four long narrative comprehension datasets. For fair comparison, all methods use GPT-4o-mini as the LLM backbone, and all non-naive RAG methods use BGE-M3 for retrieval (details in Section 3). We highlight the **best** and <u>second-best</u> results. ComoRAG is shown consistently outperform all baselines across all datasets.

featuring both question answering through free generation (QA), and multi-choice questions by selecting the best option (MC).

- **NarrativeQA** (Kociský et al. 2017): a QA dataset consisting of books and movie scripts. For ease of computation, we follow prior works and randomly sample 500 questions from the test set, with average context length 58k tokens.
- **EN.QA** from ∞BENCH (Zhang et al. 2024): a QA dataset with 351 questions on classic novels, with average context length over 200k tokens.
- **EN.MC** from ∞BENCH: a MC dataset with 229 questions on classic novels of similar length as EN.QA.
- **DetectiveQA** (Xu et al. 2024b): a MC dataset consisting of detective fiction with average length over 100k tokens. We randomly sample 20% of all stories to reduce the computational cost.

For evaluation metrics, we report both F1 and Exact Match (EM) scores for QA datasets, and report Accuracy (ACC) for MC datasets. To ensure fairness in resolving multiple-choice questions, we only expose the options during `Try-Answer`, such that no retrieval-related actions can utilize potential hints present in the options.

Baselines We employ four types of baselines as follows, covering different paradigms for long context QA.

- **LLM**: the non-RAG setting, where the entire context (capped by length 128k) is provided to the LLM directly.
- **Naive RAG**: the standard RAG setting that splits the raw context by chunks for retrieval. We set the max chunk length as 512 tokens in all experiments.
- **Enhanced RAG**: RAG methods with augmented retrieval index, including RAPTOR (Sarthi et al. 2024) that constructs a semantic summary tree over text chunks, and HippoRAGv2 (Gutiérrez et al. 2025) that builds the knowledge base for entities in text chunks. We also experimented with GraphRAG (Edge et al. 2025); however, it requires exponential computational cost for building the retrieval index, being less practical for full evaluation. We separately report GraphRAG on a subset in our extended paper version.
- **Multi-step RAG**: RAG methods with multi-step or iterative retrieval strategies. IRCoT (Trivedi et al. 2023) leverages Chain-of-Thought (CoT) as intermediate queries that iteratively retrieve evidence. Self-RAG (Asai et al. 2024) trains a dedicated critic model to control when to stop retrieval. MemoRAG (Qian et al. 2025) trains a model that compresses the global context, which generates clues as intermediate queries.

Implementation Details For the Hierarchical Knowledge Source, we follow the procedures of HippoRAGv2 and RAPTOR respectively to build the Veridical and Semantic layers; the Episodic layer employs an adaptive sliding window for narrative summaries.

For LLMs, our main experiments adopt **GPT-4o-mini** in all approaches to ensure fair comparison. We additionally tested GPT-4.1 and Qwen3-32B (Yang et al. 2025) for generalization analysis in Section 4.3. For all RAG methods, we adopt the popular model **BGE-M3** (Chen et al. 2024) for retrieval. Additionally, for naive RAG, we also experiment with larger but less practical embedding models, including NV-Embed-v2 (Lee et al. 2025) and Qwen3-Embed-8B (Zhang et al. 2025). The LLM context length for all RAG methods, including ComoRAG, is capped at 6k tokens.

For the Metacognitive Regulation loop, we set the framework to iterate for a maximum of 5 rounds. More implementation details are provided in the extended paper version.

4 Experimental Results

4.1 Main Results

Evaluation results of our main experiments are shown in Table 1. Remarkably, ComoRAG achieves the best performance upon all baselines across all datasets. Despite using the lightweight 0.3B BGE-M3 for retrieval, it significantly outperforms RAG with much larger 8B embedding models. Overall, ComoRAG demonstrates consistent improvement

Method	EN.MC	EN.QA	
	ACC	F1	EM
ComoRAG	**72.93**	**34.52**	**25.07**
Baselines			
HippoRAGv2	60.26	24.45	17.09
RAPTOR	57.21	26.33	19.65
Index			
w/o Veridical	51.97	22.24	15.88
w/o Semantic	64.63	30.82	22.65
w/o Episodic	64.63	31.48	21.47
Retrieval			
w/o Metacognition	62.01	26.95	18.53
w/o Regulation	55.02	27.95	20.59
w/o Both	54.15	25.64	17.35

Table 2: Ablation studies of ComoRAG.

for tackling long narrative comprehension, surpassing strong prior RAG methods of various paradigms.

Further analysis shows that ComoRAG remains robust even as document lengths increase, maintaining substantial advantages over the baseline in long-context settings, with the accuracy gap peaking at +24.6% for documents exceeding 150k tokens comparing with HippoRAGv2, which highlights the importance of stateful multi-step reasoning for query resolution over long and coherent contexts.

4.2 Ablation Studies

We perform ablation studies on EN.MC and EN.QA datasets by systematically removing key modules in ComoRAG. The results are shown in Table 2.

Hierarchical Knowledge Source All three knowledge layers contribute supplementary enhancements to the final performance, with the Veridical Layer being the most significant retrieval index. It provides the basis for factual-grounded reasoning, as confirmed by the 30% relative performance drop upon its removal.

Metacognition Removing the Metacognition process essentially disables the memory workspace, where all agents operate on retrieved evidence directly, without knowledge consolidation by *synthesized cues*. Disabling this module leads to a significant performance drop, as seen by the 22% relative decrease in F1 score on EN.QA, and an approximate 15% decrease in accuracy on EN.MC, underscoring the critical role of dynamic memory organization.

Regulation Removing the Regulation process cuts off the goal-oriented guidance, such that each cycle uses the same initial query for new evidence retrieval (duplicated evidence is removed), without generating probing queries that are crucial to new evidence acquisition. Disabling this module severely impacts retrieval efficiency, causing a 24% drop in accuracy on EN.MC and a 19% drop in F1 score on EN.QA.

Notably, removing both Metacognition and Regulation further degrades performance, effectively reducing the system to a one-shot resolver without multi-step reasoning.

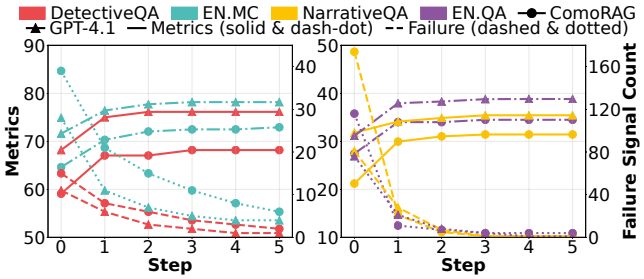

Figure 3: Performance gains from iterative probing. *GPT-4.1* marks the evaluation by using the stronger GPT-4.1 as LLM agents in ComoRAG (as opposed to GPT-4o-mini).

Overall, the ablation study results corroborate that the enhancement offered by ComoRAG stems from the synergy between its memory consolidation and dynamic evidence exploration, facilitated by the hierarchical knowledge index to provide enriched semantic information. Removing any of the core components would significantly weaken its narrative reasoning capabilities.

4.3 In-Depth Analysis of Iterative Retrieval

To further investigate the source of ComoRAG's effectiveness, this section presents a quantitative analysis of its core iterative retrieval process.

Source of Gains: From Static Bottleneck to Dynamic Reasoning Our analysis suggests that the stateful multi-step reasoning enabled by the Metacognitive loop is the key factor driving the observed improvement.

We first identify a "static bottleneck": after the initial retrieval using the original query at step 0, the single-step evaluation score shows no significant advantage over strong baselines, with less than 1% compared to the best baseline HippoRAGv2+IRCoT. However, upon activating the cognitive loop, there presents a sustained and significant improvement, raising the accuracy to 72.93% on EN.MC, as shown in Figure 3. This further supports the findings from the ablation studies, which demonstrate a significant performance drop upon removing the entire loop. Additionally, Figure 3 illustrates that the majority of the improvement occurs within 2-3 cycles, confirming the efficiency of the process. The few remaining unresolved queries are tied to the inherent reasoning limitation of the base LLM, where our next analysis shows that the ceiling performance of ComoRAG can be lifted by more capable LLMs.

Model-agnostic Generalization ComoRAG demonstrates generalization with different LLM backbones, with stronger LLMs further enhancing the reasoning process and final query resolution. To validate this, we replace GPT-4o-mini with GPT-4.1 and Qwen3-32B in the Metacognitive loop, using the same knowledge source for retrieval. The results, presented in Figure 3 and the upper section of Table 3, show a notable improvement particularly with GPT-4.1, boosting the F1 score on EN.QA from 34.52 to 38.82, and increases the accuracy on EN.MC from 72.93 to 78.17. These results demonstrate that ComoRAG effectively

Method	NarQA	EN.QA	EN.MC	DetQA
	F1	F1	ACC	ACC
ComoRAG	31.43	34.52	72.93	68.18
w/ Qwen3-32B	32.17	35.29	74.24	69.32
w/ GPT-4.1	35.43	38.82	78.17	76.14
HippoRAGv2	23.12	24.45	60.26	56.81
+ Our Loop	**29.12**	**31.76**	**68.56**	**63.64**
RAPTOR	27.84	26.33	57.21	57.95
+ Our Loop	**30.55**	**34.31**	**69.00**	**62.50**

Table 3: Efficacy of ComoRAG on model-agnostic generalization and Plug-and-Play flexibility.

leverages and unleashes the model's capabilities during its stateful iterative reasoning process.

Plug-and-Play: Flexibility To examine the modularity of our framework, we conduct further experiments by applying the Metacognitive loop of ComoRAG on existing RAG methods. As shown in the bottom section of Table 3, the cognitive loop can be seamlessly integrated with different RAG index including HippoRAGv2 and RAPTOR. This integration consistently results in significant performance improvements across all benchmarks, with accuracy on EN.MC increasing by over 8% for HippoRAGv2 and nearly 12% for RAPTOR (a similar trend is observed on EN.QA). These results demonstrate that ComoRAG could serve as a robust and flexible plug-and-play solution to enhance query resolution of existing RAG methods.

4.4 In-Depth Analysis of Query Resolution

To deepen the understanding of narrative query resolution, we roughly categorize all questions in our experimented datasets into three query types: **factoid**, **narrative**, and **inferential**, described as follows.

- **Factoid Queries**: queries answerable by a single, specific piece of information, often knowledge-seeking, e.g., *"What religion is Octavio Amber?"*
- **Narrative Queries**: queries that require an understanding of plot progression as a coherent background context, e.g., *"Where does Trace choose to live at the end of the novel?"*
- **Inferential Queries**: queries demanding reasoning beyond the literal text to understand implicit motivations, e.g., *"Why does Nils first visit Aiden's apartment?"*

To systematically investigate the dynamics of ComoRAG reasoning, we first pose the question: **what is the bottleneck in long-narrative reasoning for existing RAG methods?** Figure 4 pictures a clear diagnosis. While one-shot retrieval suffices for factoid queries, which account for over 60% of initial solution, our iterative cognitive loop is essential for resolving complex narrative queries involving global context comprehension and deeper reasoning. These constitute nearly 50% of the problems that are solved exclusively through the Metacognitive loop.

This leads to the second question: how does our framework's performance on this specific bottleneck compared to strong baselines? Figure 5 demonstrates that our method's

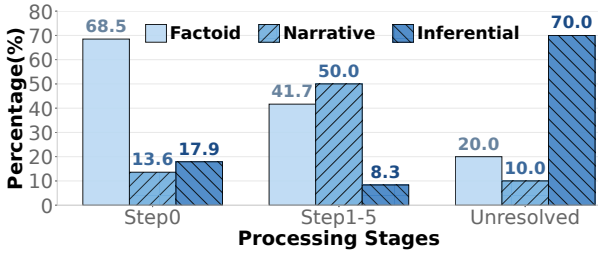

Figure 4: Distribution of solved question types.

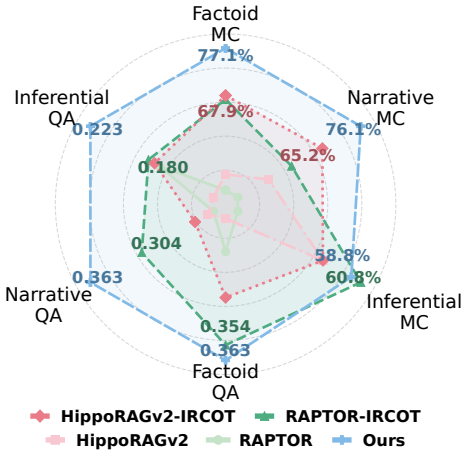

Figure 5: Benchmarking RAG methods across query types.

advantage is the most pronounced precisely in this area. On narrative queries, ComoRAG substantially outperforms the strongest baselines, achieving a **19%** relative F1 improvement on EN.QA and a **16%** accuracy gain on EN.MC. Qualitatively, Figure 2 illustrates the dynamic reasoning mechanism with the query q_{init}: *"what is the sudden purpose of Mrs. MacIntyre buying ink?"* A standard, single-step retrieval would fail on this query, as it would only find a vague clue about a "cut out newspaper", which is insufficient to form an answer. In contrast, ComoRAG initiates an iterative reasoning process by dynamically probing new queries towards a **complete evidence chain to deduce the final answer**: Mrs. McGinty recognized a photo, wanted to sell the story, and intended to write to the newspaper. We provide full reasoning details in our extended paper version.

5 Conclusion

In this work, we propose ComoRAG for long narrative reasoning, aiming to address the "stateless" limitation of conventional RAG, inspired by the human brain's Prefrontal Cortex in utilizing memory consolidation. Through our designed dynamic memory workspace and iterative probes, ComoRAG is validated on four long narrative comprehension tasks, showing that it excels at complex narrative and inferential queries where conventional stateless RAG methods fall short, marking a paradigm shift from the rather ad hoc information retrieval to cognitive reasoning towards deeper long context comprehension.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China (62372187).

References

Asai, A.; Wu, Z.; Wang, Y.; Sil, A.; and Hajishirzi, H. 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In *The Twelfth International Conference on Learning Representations*.

Chen, J.; Xiao, S.; Zhang, P.; Luo, K.; Lian, D.; and Liu, Z. 2024. M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 2318–2335. Bangkok, Thailand: Association for Computational Linguistics.

Dobbins, I. G.; and Han, S. 2006. Cue- versus Probe-dependent Prefrontal Cortex Activity during Contextual Remembering. *Journal of Cognitive Neuroscience*, 18(9): 1439–1452.

Edge, D.; Trinh, H.; Cheng, N.; Bradley, J.; Chao, A.; Mody, A.; Truitt, S.; Metropolitansky, D.; Ness, R. O.; and Larson, J. 2025. From Local to Global: A Graph RAG Approach to Query-Focused Summarization. arXiv:2404.16130.

Eisenschlos, J. M.; Yogatama, D.; and Al-Rfou, R. 2023. Needle In A Haystack: Where Is It? Finding Factual Associations in Long Texts. *arXiv preprint arXiv:2307.09288*.

Fernandez-Duque, D.; Baird, J. A.; and Posner, M. I. 2000. Executive Attention and Metacognitive Regulation. *Consciousness and Cognition*, 9(2): 288–307.

Gutiérrez, B. J.; Shu, Y.; Qi, W.; Zhou, S.; and Su, Y. 2025. From RAG to Memory: Non-Parametric Continual Learning for Large Language Models. In *Forty-second International Conference on Machine Learning*.

Jimenez Gutierrez, B.; Shu, Y.; Gu, Y.; Yasunaga, M.; and Su, Y. 2024. Hipporag: Neurobiologically inspired long-term memory for large language models. *Advances in Neural Information Processing Systems*, 37: 59532–59569.

Johnson-Laird, P. N. 1983. *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Cambridge, MA: Harvard University Press.

Kociský, T.; Schwarz, J.; Blunsom, P.; Dyer, C.; Hermann, K. M.; Melis, G.; and Grefenstette, E. 2017. The NarrativeQA Reading Comprehension Challenge. *Transactions of the Association for Computational Linguistics*, 6: 317–328.

Lee, C.; Roy, R.; Xu, M.; Raiman, J.; Shoeybi, M.; Catanzaro, B.; and Ping, W. 2025. NV-Embed: Improved Techniques for Training LLMs as Generalist Embedding Models. In *The Thirteenth International Conference on Learning Representations*.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; Riedel, S.; and Kiela, D. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 9459–9474. Curran Associates, Inc.

Liu, N. F.; Lin, K.; Hewitt, J.; Paranjape, A.; Bevilacqua, M.; Petroni, F.; and Liang, P. 2024. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12: 157–173.

Miller, J. A.; and Constantinidis, C. 2024. Timescales of learning in prefrontal cortex. *Nature Reviews Neuroscience*, 25(9): 597–610.

Qian, H.; Liu, Z.; Zhang, P.; Mao, K.; Lian, D.; Dou, Z.; and Huang, T. 2025. Memorag: Boosting long context processing with global memory-enhanced retrieval augmentation. In *Proceedings of the ACM on Web Conference 2025*, 2366–2377.

Sarthi, P.; Abdullah, S.; Tuli, A.; Khanna, S.; Goldie, A.; and Manning, C. 2024. RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval. In *International Conference on Learning Representations (ICLR)*.

Trivedi, H.; Balasubramanian, N.; Khot, T.; and Sabharwal, A. 2023. Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 10014–10037. Toronto, Canada: Association for Computational Linguistics.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; brian ichter; Xia, F.; Chi, E. H.; Le, Q. V.; and Zhou, D. 2022. Chain of Thought Prompting Elicits Reasoning in Large Language Models. In Oh, A. H.; Agarwal, A.; Belgrave, D.; and Cho, K., eds., *Advances in Neural Information Processing Systems*.

Wu, J.; Li, J.; Li, Y.; Liu, L.; Xu, L.; Li, J.; Yeung, D.-Y.; Zhou, J.; and Yu, M. 2025. SitEmb-v1.5: Improved Context-Aware Dense Retrieval for Semantic Association and Long Story Comprehension. arXiv:2508.01959.

Xu, L.; Li, J.; Yu, M.; and Zhou, J. 2024a. Fine-Grained Modeling of Narrative Context: A Coherence Perspective via Retrospective Questions. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5822–5838. Bangkok, Thailand: Association for Computational Linguistics.

Xu, Z.; Ye, J.; Liu, X.; Sun, T.; Liu, X.; Guo, Q.; Li, L.; Liu, Q.; Huang, X.; and Qiu, X. 2024b. DetectiveQA: Evaluating Long-Context Reasoning on Detective Novels. *ArXiv*, abs/2409.02465.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W.; Salakhutdinov, R.; and Manning, C. D. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In Riloff, E.; Chiang, D.; Hockenmaier, J.; and Tsujii, J., eds., *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2369–2380. Brussels, Belgium: Association for Computational Linguistics.

Zhang, X.; Chen, Y.; Hu, S.; Xu, Z.; Chen, J.; Hao, M.; Han, X.; Thai, Z.; Wang, S.; Liu, Z.; et al. 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)*, 15262–15277.

Zhang, Y.; Li, M.; Long, D.; Zhang, X.; Lin, H.; Yang, B.; Xie, P.; Yang, A.; Liu, D.; Lin, J.; Huang, F.; and Zhou, J. 2025. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. arXiv:2506.05176.