/TemplateVersion (2026.1)

ComoRAG: 一个认知启发式记忆组织的 RAG，用于有状态的长叙述推理

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]Pattern Recognition Center, WeChat AI, Tencent

Abstract

对长篇故事和小说的叙事理解一直是一个具有挑战性的领域，这归因于其复杂的情节线索以及角色和实体之间错综复杂且往往不断演变的关系。鉴于 LLM 在长时间背景下推理能力的下降以及高昂的计算成本，基于检索的方法在实践中仍占据重要地位。然而，传统 RAG 方法可能会因其无状态的单步检索过程而显得不足，这种过程往往忽视了在长距离背景中捕捉相互关联关系的动态特性。在这项工作中，我们提出了 ComoRAG，它的核心原则是叙事推理不是一次性过程，而是新证据获取与过去知识整合之间动态、不断发展的相互作用，类似于人类在用与记忆相关的信号进行脑内推理时的认知过程。具体来说，当遇到推理困境时，ComoRAG 进行迭代推理循环，同时与动态记忆工作空间互动。在每个循环中，它生成探测性查询以设计新的探索路径，然后将检索到的新方面证据整合到一个全局记忆池中，从而支持查询解决时形成连贯的背景。在四个具有挑战性的长文本叙事基准（超过 20 万令牌）中，ComoRAG 相较于最强基线持续取得最高达 11 % 的相对优势。进一步分析显示，ComoRAG 尤其适用于需要全球理解的复杂查询，为基于检索的长文本理解提供了一个原则性、认知驱动的范式以实现有状态推理。我们的代码已公开发布在 https://github.com/EternityJune25/ComoRAG 。

1 介绍

长篇叙事理解的核心挑战不仅仅在于连接离散的证据片段，这项任务更自然地被定义为多跳问答（QA），而在于动态地进行认知综合，以把握必要的背景和内容进展 (Xu et al. 2024a) 。与多跳 QA (Yang et al. 2018) 不同的是，后者寻求通过固定事实的静态路径，而叙事理解需要模拟人类读者：不断构建和修正关于情节、角色及其动机变化的整体心智模型 (Johnson-Laird 1983) 。这一过程的复杂性由经典的叙事问题“为什么斯内普杀了邓布利多?”很好地例证了。这需要从跨越多本书的不同线索中编织完整的证据网——邓布利多的绝症、不可破誓言，以及斯内普深藏不露的忠诚。这些线索的真正意义只有在事后才得以完全理解。这种能力我们称之为状态推理：它不仅要求连接静态证据；更要求维护一个不断随新信息出现而更新的动态记忆特征于叙事 。近年来，长文本上下文的 LLMs 在“Needle in a Haystack”测试 (Eisenschlos, Yogatama, and Al-Rfou 2023) 中展示了有前途的表现。然而，这些模型处理长篇叙事（超过 20 万标记）的能力仍然受到有限上下文窗口的限制。此外，随着输入长度增加，这些模型易患“中间迷失”问题 (Liu et al. 2024) ，而这会增加困惑性并削弱生成质量。这一限制在需要状态推理的叙事理解任务中尤为明显。为此，检索增强生成（RAG）(Lewis et al. 2020) 已成为用 LLMs 处理长文本理解的重要策略。

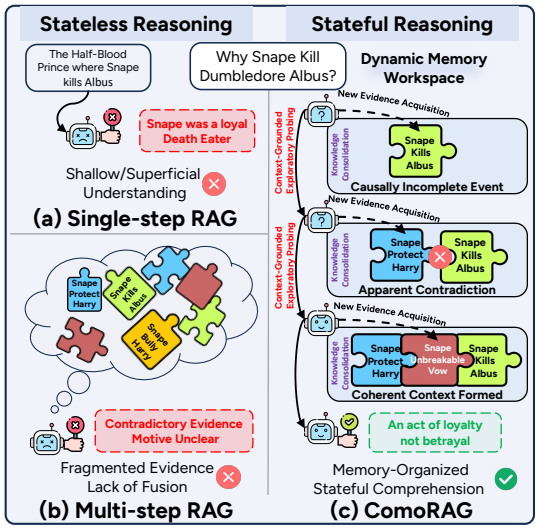

Figure 1: RAG 推理范式的比较。

然而，现有的 RAG 方法仍然难以有效解决这一挑战。高级单步检索因其静态索引而受到限制。这包括诸如 RAPTOR (Sarthi et al. 2024) 之类的方法，该方法对文本块进行聚类和摘要，以便在不同的细节层次上进行检索；HippoRAGv2 (Gutiérrez et al. 2025) 则通过构建知识图谱来模拟人类海马体，以在单次检索步骤中实现多跳推理。然而，单步方法依赖于一次性静态检索，可能导致浅层理解。例如，在图 1 (a) 中，有关斯内普的证据可能会误导模型做出错误推断。

作为一种补救措施，多步检索方法提供了一个更有前途的方向，例如将检索过程与链式推理相结合的 IR-CoT (Trivedi et al. 2023) ，以及训练模型以自适应检索和反思证据的 Self-RAG (Asai et al. 2024) ，以及通过使用双系统架构从压缩的全局上下文中生成线索的

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

MemoRAG (Qian et al. 2025) 。这些方法都旨在通过迭代检索获得更丰富的上下文。然而，它们的检索步骤通常是独立的，缺乏在显式叙述推进过程中的连贯推理，特点是碎片化的证据和无状态的理解。如图 1 (b) 所示，由于缺乏动态记忆，多步检索无法整合矛盾的证据，比如“斯内普保护/欺负哈利”，不能理解他行动的演变，最终无法得出正确的答案。

在这项工作中，我们从人脑中前额叶皮层（PFC）的功能中寻求灵感，PFC 采用一种称为元认知调节的复杂推理过程 (Fernandez-Duque, Baird, and Posner 2000) 。这个过程并不是单一的动作，而是新证据获取与随后的知识整合之间的动态互动，由目标导向的记忆探测驱动 (Dobbins and Han 2006; Miller and Constantinidis 2024) 。在整合过程中，新发现会与过去的信息相结合，以构建一个不断发展的连贯叙述。这个迭代循环使 PFC 能够持续评估其理解并调整其策略，为我们的框架提供一种具有状态保持特性的推理方法的直接认知蓝图。

我们引入了 ComoRAG ，这是一种由 co 启发、由 m 内存组织的 RAG 框架，模仿人类前额叶皮层（PFC）以实现真正的状态化推理。其核心是一个在内存工作空间上运行的动态认知循环，该循环积极探索并整合新证据以构建连贯的叙述理解。

这一过程，如图 1 (c) 所示，是一个演变推理状态的闭环。面对一个复杂的问题，比如“为什么斯内普会杀邓布利多?”，系统的记忆状态从一个初始的“因果不完整事件”（斯内普杀死阿不思），演变为由于发现矛盾信息（斯内普保护哈利）而产生的“明显矛盾”，并最终通过更深入的探索和证据融合形成逻辑一致的连贯背景。只有在这个最终的、完整的认知状态下，ComoRAG 才能执行正确的状态推理，得出“这是一种忠诚的行为，而非背叛”的深刻见解。

这种认知启发的设计在四个具有挑战性的长文本叙事基准中实现了显著的提升。ComoRAG 在每个数据集上始终超越各类强力基线。我们的分析揭示了几个关键发现。首先，这些收益直接源于认知循环，它将静态知识库转变为动态推理引擎；例如，EN.MC 的准确率从静态检索基线的 64.6 % 跳升至 72.9 %，并且性能在大约 2-3 次循环中高效收敛。其次，我们的框架在需要整体理解故事情节发展的叙述查询上表现优异，在这些具有挑战性的问题类型上实现了高达 19 % 的相对 F1 提升，而其他方法却未能做到。最后，我们的框架展示了显著的模块化和普适性。其核心循环可以灵活地集成到现有的 RAG 方法中，如 RAPTOR，其直接带来了 21 % 的相对准确性提升。此外，切换到一个更强大的模型作为骨干 LLM 代理可以提升整个认知循环中的推理能力，将准确率从 72.93 % 提升到 78.17 %。这些结果共同验证了 ComoRAG 提供了一种基于检索的长文本叙事理解的新范式，面向具有状态性的推理。

2 方法论

我们引入了 ComoRAG，这是一种自主认知架构，旨在形式化和实现如引言中所述的元认知调节过程。该架构的设计直接受前额皮质（PFC）功能机制的启发，并建立在三个概念支柱上：（1）用于深入上下文理解的分层知识源；（2）用于跟踪和整合多轮推理的动态记忆工作空间；以及（3）驱动整个解决过程的元认知控制循环。

2.1 问题表述：走向有原则的叙事推理

我们的目标是设计一个框架，用于在 RAG 场景中进行有状态推理。尤其是，它旨在首先解决那些需要理解全局上下文的查询，这在叙述中很常见，而传统的 RAG 可能无法基于查询的表面形式识别相关上下文。形式上，设初始查询为 q_{init} ，知识来源 $\mathcal{X}$ 是基于原始上下文派生的，我们的框架 F 通过一系列自适应操作来在离散时间步骤 $t = 1, \ldots, T$ 中通过底层记忆控制得出最终答案 A_{final} 。

在每一步的开始，t ，F 确定其推理的重点——一组新的探测查询 $\mathcal{P}^{(t)}$ ，代表需要寻找的新的信息，这些信息可能逻辑上深化查询理解并最终补充答案解析。通过 $\mathcal{P}^{(t)}$ 在每一步检索到的新信息，该框架利用到前一步维持的全局内存池 $\mathcal{M}_{pool}^{(t-1)}$ ，并产生最终答案或失败信号，指示推理僵局——并将内存池更新到 $\mathcal{M}_{pool}^{(t)}$ ，完成一个在知识源、记忆空间和检索操作之间协同的认知循环。

2.2 层级知识源

为了克服单一表示法所带来的限制，我们的框架首先构建一个层次知识索引 $\mathcal{X}$ ，用于检索，这个索引用三个互补的认知维度来建模原始文本，类似于前额叶皮层（PFC）如何整合来自大脑不同区域的不同记忆类型，特别是支持从原始证据到抽象关系的跨层推理。

为了确保所有推理都可以追溯到源证据，首先建立一个真实层 $\mathcal{X}^{ver}$ ，由原始文本块直接构成，类似于人类记忆中精确回忆事实细节。为了更准确地检索文本块，我们指导大型语言模型为每个文本块生成知识三元组（主体 - 谓词 - 宾语）。这些三元组参与每次检索，并增强传入查询与相应文本块之间的匹配，这已被 HippoRAG (Jimenez Gutierrez et al. 2024) 证实有效。进一步的细节在附录 B 中描述。

语义层：抽象主题结构。 为了捕捉超越长距离上下文依赖性的主题和概念联系，建立了一个语义层 $\mathcal{X}^{sem}$ ，这一构建受到先前工作 RAPTOR 的启发，该工作采用 GMM 驱动的聚类算法递归地将语义相似的文本块总结为一个分层汇总树。我们认为这种语义抽象对更深入的理解是必要的，并遵循相同的公式化。这些总结节点使得该框架能够检索超越表面层次的概念信息。

情节层：重构叙述流程。 前两层提供了事实细节和高层次概念的视角。然而，它们缺乏时间发展或情节推进，这对叙事来说可能尤其重要。为了实现具有长程因果链的这种视角，我们引入了情节层 $\mathcal{X}^{epi}$ ，其旨在通过捕捉顺序叙事发展来重建情节线和故事弧。该过程的特点是跨文本块的滑动窗口摘要；然后，每个生成的节点都是根据时间线汇总连续或因果相关事件的叙事发展的摘要。可以选择递归地应用滑动窗口过程，以形成更高层次的内容推进视图，作为知识来源的一部分提取不同层次的叙事流。

2.3 元认知调节的架构

ComoRAG 的核心是一个控制循环，完全实现了元认知调节的概念。它由一个用于反思和规划的调节过程和一个用于执行推理和内存管理的元认知过程组成，后者利用记忆工作区进行操作。

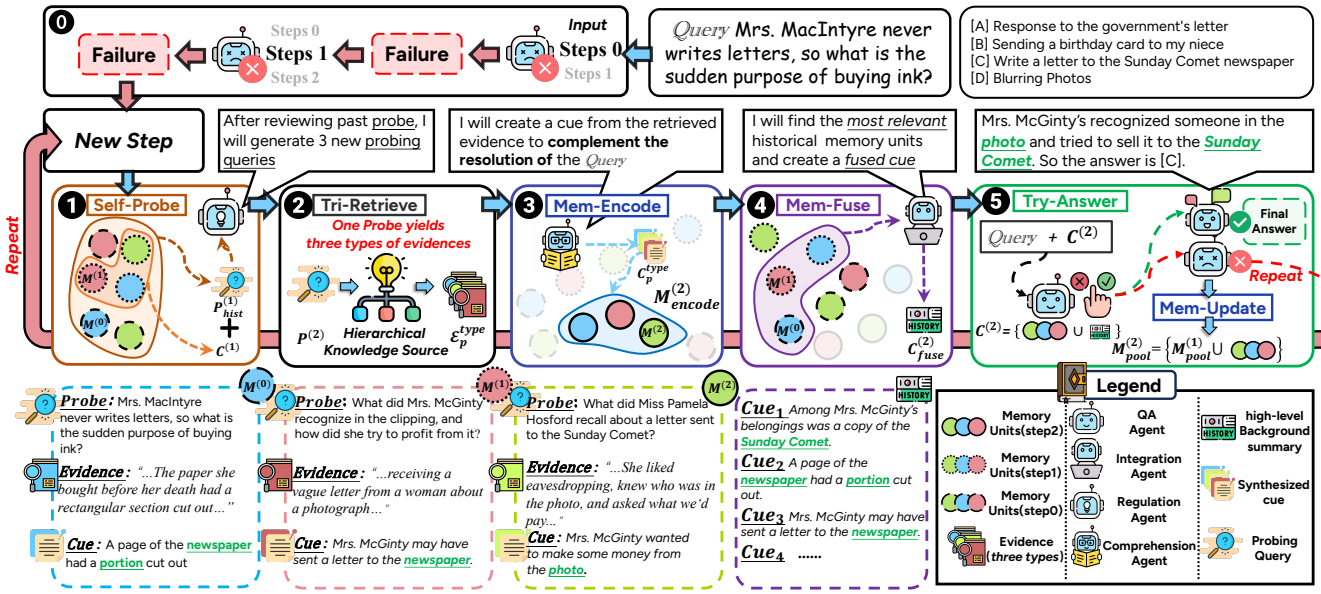

Figure 2: 对 ComoRAG 的说明。受到推理僵局（失败）的触发，元认知调节环包括了在第 2.3 节中描述的五个核心操作：1） Self-Probe ，根据过去的记忆单元设计新的探索性探测查询；2） Tri-Retrieve ，从三个知识来源中检索证据；3） Mem-Encode ，形成关于新证据如何补充最终查询解决的新记忆单元；4） Mem-Fuse ，生成提示以整合新旧记忆单元；5） Try-Answer ，利用这一周期中产生的新记忆信息进行问题解答。

动态存储工作区 记忆工作区包含记忆单元，通过元认知调节充当连贯的多步骤探索和推理的桥梁。每个记忆单元 m 功能性地结束一个检索操作，该操作表示为三个元素的元组：$m = (p, \mathcal{E}_p^{type}, \mathcal{C}_p^{type})$ ，其中 p 是触发此检索的探测查询；$\mathcal{E}_p^{type}$ 是从单个知识层（$type \in \{ver, sem, epi\}$ ）检索到的同质证据集合；而 $\mathcal{C}_p^{type}$ 是一个合成线索，反映了这些由探测 p 检索到的证据如何补充原始查询 q_{init} 的理解和解析。具体来说，$\mathcal{C}_p^{type}$ 是由大语言模型（担任理解代理的角色）生成的，记为 $\mathcal{C}_p^{type} = \pi_{cue}(q_{init}, p, \mathcal{E}_p^{type})$ 。

通过每次检索形成一个记忆单元 $(p, \mathcal{E}_p^{type}, \mathcal{C}_p^{type})$ 被定义为 Mem-Encode 操作。在下面描述的推理循环中，记忆工作区/池将被利用和更新。

如果前一个推理循环 $t-1$ 以失败告终，那么在下一个推理循环/步骤 t 开始时会调用监管过程。核心操作 Self-Probe 计划新的探测查询，这些查询获取的信息可能有助于最终答案，从而设计出新的探索路径以打破僵局。这是由一个名为 π_{probe} 的监管代理协调的，其决策是基于对先前失败的反思，探索更多必要的背景或相关信息，以全面理解并解决原始查询。 Self-Probe 接受三个输入：(1) 终极目标 q_{init} ；(2) 到上一步结束时的完整探索探测历史 $\mathcal{P}_{hist}^{(t-1)}$ ；以及 (3) 导致失败的直接知识缺口，由前一步生成的所有记忆单元的综合线索具体化，记作 $\{\mathcal{C}\}^{(t-1)}$ 。其输出 $\mathcal{P}^{(t)}$ 是一个新的、战略性的检索探针集合，用于当前循环 t ：

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

元认知过程 元认知过程接收此循环的新探针 $\mathcal{P}^{(t)}$ ，并执行推理以解决原始查询，同时跟踪内存空间的进展。它包括一系列操作，详细描述如下。

Tri-Retrieve ：对于每一个探测查询 $p \in \mathcal{P}^{(t)}$ ，在每个知识层 $\mathcal{X}^{type}$ 上进行检索，条件为 $type \in \{ver, sem, epi\}$ ，这样就可以在一个标准的密集段落检索范式中，检索出与 p 每一层具有高嵌入相似性的证据，每条证据可以是原始文本块、语义聚类摘要或叙述流程摘要。

Mem-Encode ：对于每个 p 和 $type$ ，检索到的证据会立即被上述的 Mem-Encode 处理，以生成一个新的记忆单元，该单元跟踪该特定探测如何补充到最终答案。这一步生成的所有记忆单元的数量可以用 $|\mathcal{M}_{encode}^{(t)}| = 3 \times |\mathcal{P}^{(t)}|$ 表示。

Mem-Fuse ：以上步骤中的新记忆单元 $\mathcal{M}_{encode}^{(t)}$ 主要强调当前循环中探测的方面。为了充分利用过去的经验和历史知识，框架进一步从现有记忆池 $\mathcal{M}_{pool}^{t-1}$ 中的过去单元中识别相关的合成线索，然后生成一个新的合成线索以融合过去的相关证据。设 $\mathcal{M}_{pool}^{t-1} \circ q_{init}$ 表示其线索与 q_{init} 有高嵌入相似度的过去记忆单元，并将一个大型语言模型表示为融合代理 π_{fuse} ，该模型将这些相关的过去证据合成为一个高级背景摘要，则融合过去记忆的新线索 $\mathcal{C}_{fuse}^{(t)}$ 为：

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

Try-Answer ：利用 $\mathcal{M}_{encode}^{(t)}$ 中的新探测证据和 $\mathcal{C}_{fuse}^{(t)}$ 的过去融合线索，应用一个 QA 代理 π_{QA} 于这些上下文中，以生成此循环的最终输出 $O^{(t)}$ ：

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

具体来说，指示一个 LLM 将这些最新的证据和过去的背景作为上下文，并确定原始查询是否可以解决。它或者给出最终答案并终止整个推理循环，或者发出失败信号并继续进行到下一步。

Mem-Update ：在一个循环的最后一步中，这只是将新生成的记忆单元纳入全局池中，并对其嵌入进行编码，以便将来的检索和推理：

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

通过从 Tri-Retrieve 到 Mem-Update 的上述六个步骤，实现了认知循环的一个周期。在 $t = 0$ 中的初始步骤中，ComoRAG 从 Tri-Retrieve 的一轮开始，接着是 Try-Answer 。如果信号显示失败，它将启动探索路径上的有状态推理的元认知循环，其特征在于与记忆工作区的联锁操作，从而能够应对复杂的叙事理解。

从本质上讲，我们的框架把握住了一个原则，即对于长篇上下文理解，特别是在整个语境通过故事情节进展 (Xu et al. 2024a) 紧密关联的叙述中，查询解析不是一个线性流程；而是一个新的证据获取和过去知识整合之间动态、不断演变的互动，类似于人类的认知过程。整个过程的进一步说明详见附录 A 的算法；每个 LLM 代理所使用的详细提示见附录 D 。

3 实验设置

数据集 我们的实验涵盖了四个长篇语境叙事理解的数据集，以进行全面评估，特点是通过自由生成进行问答 (QA)，以及通过选择最佳选项进行的多选题 (MC)。

- NarrativeQA (Kociský et al. 2017) ：一个由书籍和电影剧本组成的问答数据集。为了简化计算，我们遵循之前的工作，并从测试集中随机抽取 500 个问题，平均上下文长度为 58k 标记。
- 来自 ∞ BENCH (Zhang et al. 2024) 的 EN.QA：一个包含 351 个经典小说问题的 QA 数据集，平均上下文长度超过 20 万标记。
- 来自 ∞ BENCH 的 EN.MC：一个包含 229 个关于经典小说问题的 MC 数据集，与 EN.QA 长度相似。
- DetectiveQA (Xu et al. 2024b) ：一个侦探小说的多选数据集，平均长度超过 10 万标记。我们随机抽取所有故事的 20 %，来降低计算成本。

针对评估指标，我们报告问答数据集的 F1 和完全匹配 (EM) 得分，并针对多项选择数据集报告准确率 (ACC) 。为了确保多项选择题的解答公平，我们仅在 Try-Answer 期间展示选项，以避免检索相关行为利用选项中可能存在的提示。

基线 我们采用以下四种基线，涵盖了长文档问答的不同范式。

- LLM：非 RAG 设置，其中整个上下文 (长度上限为 128k) 直接提供给 LLM。
- Naive RAG：标准的 RAG 设置，通过块来分割原始上下文以进行检索。在所有实验中，我们将最大块长度设置为 512 个标记。
- 增强型 RAG：通过增强的检索索引进行的 RAG 方法，包括 RAPTOR (Sarthi et al. 2024) ，它在文本块上构建语义摘要树，以及 HippoRAGv2 (Gutiérrez et al. 2025) ，它为文本块中的实体构建知识库。我们还尝试了 GraphRAG (Edge et al. 2025) ；然而，它在构建检索索引时需要指数级的计算成本，因此对于全面评估而言不太实用。我们在附录 B 中的一个子集上单独报告 GraphRAG。
- 多步 RAG：具有多步骤或迭代检索策略的 RAG 方法。IRCoT (Trivedi et al. 2023) 利用链式思维 (CoT) 作为中间查询来迭代检索证据。Self-RAG (Asai et al. 2024) 训练一个专门的评判模型来控制何时停止检索。MemoRAG (Qian et al. 2025) 训练一个模型来压缩全局上下文，从而生成线索作为中间查询。

实现细节 对于分层知识源，我们分别遵循 HippoRAGv2 和 RAPTOR 的程序来构建真实和语义层；情节层使用适应性滑动窗口进行叙述摘要，具体描述见附录 B 。

对于 LLM，我们的主要实验在所有方法中均采用 GPT-4o-mini 以确保公平比较。我们还在第 4.2 节中测试了 GPT-4.1 和 Qwen3-32B (Yang et al. 2025) 以进行泛化分析。对于所有 RAG 方法，我们采用流行模型 BGE-M3 (Chen et al. 2024) 进行检索。此外，对于简单的 RAG，我们还试验了更大但实用性较低的嵌入模型，包括 NV-Embed-v2 (Lee et al. 2025) 和 Qwen3-Embed-8B (Zhang et al. 2025) 。对于所有 RAG 方法，包括 ComoRAG ，LLM 的上下文长度限制为 6k tokens。

对于元认知调节循环，我们将框架设置为最多迭代 5 轮。

更多的实现细节在附录 B 中进一步提供。

4 实验结果

4.1 主要结果

我们主要实验的评估结果如表 1 所示。值得注意的是，ComoRAG 在所有基线和所有数据集上都取得了最佳性能。尽管使用了轻量级的 0.3B BGE-M3 进行检索，它显著超越了拥有更大 8B 嵌入模型的 RAG。总体而言，ComoRAG 展示了在处理长篇叙事理解方面的一贯改进，超过了各种方法范式下的强大 RAG 前辈。

经过仔细检查，ComoRAG 在两个具有超长上下文的 ∞ BENCH 数据集上展示了明显的优势。更广泛地说，图 3 显示出 ComoRAG 更加稳健且对较长的上下文不敏感，保持了其在 HippoRAGv2 上的效能，准确度差距在超过 150k 标记的文档时达到了 +24.6 %，这突显了在长且连贯的上下文中进行查询解析时状态化多步骤推理的重要性。

我们通过系统地去除 ComoRAG 中的关键模块，对 EN.MC 和 EN.QA 数据集进行消融研究。结果如表 2 所示。

层次知识来源 这三层知识对最终性能都提供了补充性的提升，其中真实性层是最重要的检索指标。它为基于事实的推理提供了基础，如去除它后大约导致 30 % 相对性能下降所证实。

移除元认知过程本质上禁用了记忆工作区，所有代理在这个环境中直接对检索到的证据进行操作，而无需通过合成线索进行知识整合。禁用该模块会导致显著的性能下降，如在 EN.QA 上 F1 分数相对下降了%_22，在 EN.MC 上的准确率大约下降了%_15，这强调了动态记忆组织的重要作用。

移除调节过程会切断目标导向的指导，这样每个循环在检索新证据时使用相同的初始查询 (重复的证据被移除)，而不会生成对新证据获取至关重要的试探性查询。禁用此模块会严重影响检索效率，导致在 EN.MC 上的准确率下降 24 %，在 EN.QA 上的 F1 分数下降 19 %。

值得注意的是，移除元认知和调控会进一步降低性能，实际上将系统简化为一个没有多步骤推理的一次性解决方案。总体而言，消融研究的结果证明，ComoRAG

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	17.80	34.34	24.57	61.13	62.50	30.76	21.19	61.82
	Qwen3-Embed-8B	24.19	15.60	25.79	17.95	65.50	61.36	24.99	16.78	63.43
Enhanced RAG	RAPTOR	27.84	17.80	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	31.35	16.00	32.09	19.36	63.76	64.77	31.72	17.68	64.27
	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: 在四个长篇叙事理解数据集上的评估结果。为了公平比较，所有方法使用 GPT-4o-mini 作为 LLM 骨干，所有非简单的 RAG 方法使用 BGE-M3 进行检索（详情参见第 3 节）。我们强调了最佳和 second-best 结果。结果显示 ComoRAG 在所有数据集上始终优于所有基线。

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: ComoRAG 的消融研究。

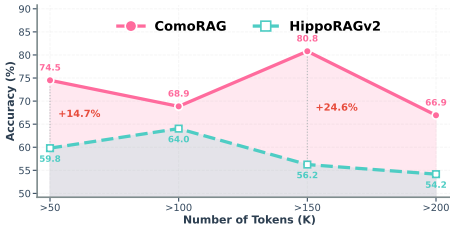

Figure 3: 在多选数据集上不同文档长度的平均准确率。ComoRAG 在较长上下文中显示出比基线更强的鲁棒性。

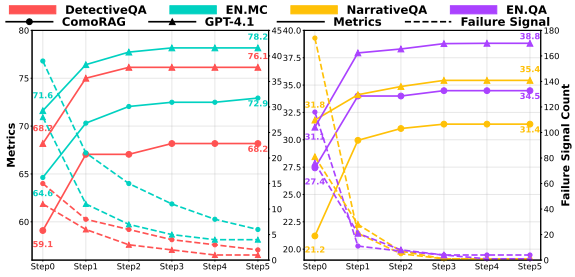

Figure 4: 迭代试探带来的性能提升。在 ComoRAG 中，GPT-4.1 作为 LLM 代理来进行评价(而不是使用 GPT-4o-mini)。

提供的增强效果源于其记忆巩固与动态证据探索之间的协同作用，这一作用通过分层知识索引提供丰富的语义信息。移除任何核心组件都会显著削弱其叙事推理能力。

4.2 迭代检索的深入分析

为了进一步研究 ComoRAG 有效性的来源，本节对其核心迭代检索过程进行了定量分析。

我们的分析表明，由元认知循环支持的有状态多步骤推理是推动观察到的改进的关键因素。

我们首先识别出一个“静态瓶颈”：在步骤 0 使用原始查询进行初始检索后，单步评估分数相比强大的基准没有显著优势，与最佳基准 HippoRAGv2+IRCoT 相比，少于 1 %。然而，在激活认知循环后，呈现出持续且显著的改进，将 EN.MC 上的准确性提升至 72.93 %，如图 4 所示。这进一步支持了来自消融研究的发现，即去除整个循环后性能显著下降。此外，图 4 显示，大部分的改进发生在 2-3 个周期内，确认了该过程的效率。剩下的几个未解决的查询与基础 LLM 固有的推理限制有关，我们的下一次分析显示，通过简单地切换到更有能力的 LLMs 可以提升 ComoRAG 的顶级表现。

模型无关的泛化 ComoRAG 展示了使用不同 LLM 主干的泛化能力，随着更强大的 LLM 进一步增强推理过程和最终查询解决。为了验证这一点，我们将 Metacognitive 环中的 GPT-4o-mini 替换为 GPT-4.1 和 Qwen3-32B，使用相同的知识源进行检索。图 4 和表 3 的上部分所展示的结果显示了显著的改进，尤其是在 GPT-4.1 上，将 EN.QA 的 F1 得分从 34.52 提升至 38.82，并将 EN.MC 的准确率从 72.93 提高到 78.17。这些结果表明 ComoRAG 在其有状态迭代推理过程中有效地利

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: ComoRAG 在模型无关泛化和即插即用灵活性方面的有效性。

用并释放了模型的能力。

为了检验我们框架的模块化，我们通过将 ComoRAG 的元认知循环应用于现有的 RAG 方法来进行进一步的实验。如表格 3 的底部部分所示，认知循环可以无缝地与不同的 RAG 索引整合，包括 HippoRAGv2 和 RAPTOR。这种整合在所有基准上持续带来显著的性能提升，EN.MC 的准确率对 HippoRAGv2 提高超过 8 %，对 RAPTOR 则接近 12 %（在 EN.QA 上也观察到类似趋势）。这些结果表明，ComoRAG 可以作为一个稳健且灵活的即插即用解决方案来增强现有 RAG 方法的查询解析能力。

4.3 查询解析的深入分析

为了加深对叙述性查询解析的理解，我们将实验数据集中的所有问题大致分为三种查询类型：事实性、叙述性和推理性，具体描述如下（详细信息见附录 C）。

- 事实型查询：通过单一具体的信息即可回答的查询，通常是寻求知识的，例如，“Octavio Amber 的宗教是什么?”
- 叙事查询：需要理解情节进展作为连贯的背景上下文的查询，例如，“Trace 最终选择在哪里居住?”
- 推理性查询：这些查询需要超过字面文本的推理以理解隐含动机，例如，“尼尔斯第一次到艾登的公寓拜访的主要原因是什么?”

为了系统地研究 ComoRAG 推理的动态，我们首先提出问题：现有 RAG 方法在长篇叙事推理中的瓶颈是什么？图 5 展示了一个清晰的诊断。虽然一次性检索足以解决占超过%_60 的初始解决方案的事实性查询，但我们的迭代认知循环对于解决涉及全球背景理解和更深入推理的复杂叙述性查询是必不可少的。这些占到近%_50 的问题，仅通过元认知循环得到解决。

这引出了第二个问题：我们的框架在这个特定瓶颈上的表现如何与强大的基线相比？图 6 显示，我们的方法的优势在这一领域表现得尤为显著。在叙事查询上，ComoRAG 显著超越了最强的基线，在 EN.QA 上实现了 19 % 的相对 F1 提升，并在 EN.MC 上获得了 16 % 的准确度提升。通过解决这些查询，我们证明了我们的框架的成功不仅仅是一个总体的改进，而是对叙事查询类型的一个有针对性的、有效的解决方案——这是实现真正叙事理解的基石，而之前基于检索的叙事推理方法对此一直面临挑战。

定性地说，图 2 展示了使用查询 q_{init} 的动态推理机制：“麦金太太从不写信，那突然买墨水的目的是什么?”对于这个查询，标准的单步检索将会失败，因为它只会找到关于“剪报”的模糊线索，这不足以形成一个答案。相比之下，ComoRAG 通过动态探测新查询和证据来启动迭代推理过程，以实现完全解决，从而构建一个完整的证据链来推导最终答案：麦金太太认出了一张照片，想要出售这个故事，并打算写信给报社。我们在附录 E 中提供了完整的推理细节。

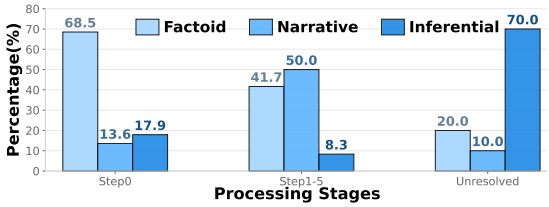

Figure 5: 解决的问题类型按处理阶段的分布。

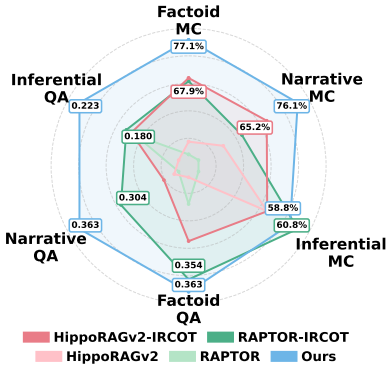

Figure 6: 在不同查询类型中对 RAG 方法进行基准测试。

5 结论

在这项工作中，我们提出了 ComoRAG 用于长篇叙述推理，旨在解决传统 RAG 的“无状态”限制。ComoRAG 特别受到人脑前额皮质的启发；通过动态记忆工作空间和迭代探测，它将分散的证据融合成一个连贯的背景，以实现叙述进程中的有状态推理。实验验证了 ComoRAG 通过在复杂叙述和推理查询中表现出色，克服了现有方法的瓶颈，标志着从信息检索到认知推理的范式转变，以实现对更深层次长背景的理解。

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 Contex-

tual 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.

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.

A ComoRAG 算法

Algorithm 1: ComoRAG （在第 2 节中描述）

Require: Initial Query q_{init} , Knowledge Source $\mathcal{X}$, Max Iterations T
Ensure: The final answer O or a failure signal

function ComoRAG($q_{init}, \mathcal{X}, T$)
  $\mathcal{M}_{pool}^{(0)}, \mathcal{P}_{hist}^{(0)}, \{\mathcal{C}\}^{(0)} \leftarrow \emptyset, \emptyset, \emptyset$ ▷ Initialize Memory Pool, Probing History, and Synthesized Cues
  $\mathcal{E}^{(0)} \leftarrow$ Tri-Retrieve $(\{q_{init}\}, \mathcal{X})$
  $O^{(0)} \leftarrow$ Try-Answer $(q_{init}, \mathcal{E}^{(0)})$
  if $O^{(0)} \neq$ FailureSignal then
    return $O^{(0)}$ ▷ Return immediately if successful
  end if
  ▷ Triggered only if initial attempt fails
  $\mathcal{M}_{encode}^{(0)} \leftarrow$ Mem-Encode $(q_{init}, \mathcal{P}_{hist}^{(0)}, \mathcal{E}^{(0)})$
  $\mathcal{M}_{pool}^{(0)} \leftarrow$ Mem-Update $(\mathcal{M}_{pool}^{(0)}, \mathcal{M}_{encode}^{(0)})$
  $\mathcal{P}_{hist}^{(0)} \leftarrow q_{init}$
  $\{\mathcal{C}\}^{(0)} \leftarrow \mathcal{M}_{pool}^{(0)}$
  for $t = 1, \ldots, T$ do
    $\mathcal{P}^{(t)} \leftarrow$ Self-Probe $(q_{init}, \mathcal{P}_{hist}^{(t-1)}, \{\mathcal{C}\}^{(t-1)})$
    $\mathcal{E}^{(t)} \leftarrow$ Tri-Retrieve $(\mathcal{P}^{(t)}, \mathcal{X})$
    $\mathcal{M}_{encode}^{(t)} \leftarrow$ Mem-Encode $(q_{init}, \mathcal{P}^{(t)}, \mathcal{E}^{(t)})$
    $\mathcal{C}_{fuse}^{(t)} \leftarrow$ Mem-Fuse $(q_{init}, \mathcal{M}_{pool}^{(t-1)} \circ q_{init})$
    $O^{(t)} \leftarrow$ Try-Answer $(q_{init}, \mathcal{M}_{encode}^{(t)}, \mathcal{C}_{fuse}^{(t)})$
    if $O^{(t)} \neq$ FailureSignal then return $O^{(t)}$
    end if
    $\mathcal{M}_{pool}^{(t)} \leftarrow$ Mem-Update $(\mathcal{M}_{pool}^{(t-1)}, \mathcal{M}_{encode}^{(t)})$
    $\mathcal{P}_{hist}^{(t)} \leftarrow \mathcal{P}_{hist}^{(t-1)} \cup \mathcal{P}^{(t)}$
    $\{\mathcal{C}\}^{(t)} \leftarrow \mathcal{M}_{pool}^{(t)}$
  end for
  return FailureSignal
end function

B 实现细节

如在第 2.2 节中描述的，ComoRAG 通过构建一个分层知识源来增强大型语言模型，其中真实层是一个由原始上下文的文本块组成的基础组件。我们主要遵循 HippoRAGv2 (Gutiérrez et al. 2025) 的构建过程，添加知识图谱（KGs）与文本块之间的映射以便于检索。为了构建知识图谱，利用大型语言模型（LLM）提取（主体-谓词-客体）的知识三元组。这些来自文档的三元组然后被聚合以形成一个统一的知识图谱。最后，一个检索优化的编码器通过识别和连接语义相似实体（同义词）为此图谱添加补充的边。因此，真实层的检索遵循 HippoRAGv2 利用知识图谱实现更准确的检索。该层的统计数据详见表 4 。

Layer	Count	NarQA	EN.QA	EN.MC	DetQA
Veridical	# of Chunks	4446	26 465	47 074	2406
	# of Entities	33 810	292 170	401 040	30 969
	# of Triples	51 012	372 339	576 595	33 696

Table 4: 跨数据集的真实层统计。

B.1 情节层

为了构建情节层，文本块序列被总结。由于上下文长度可能显著变化，为此选择滑动窗口大小进行总结时存在权衡：较大的窗口可能对简短叙述来说过于粗略，而较小的窗口可能效率低下，并且无法捕捉长篇内容的长距离依赖。因此，我们根据文档中总的文本块数量 N 动态调整窗口大小 W。具体的启发式方法如下。

- 对于中短篇叙述（$N \leq 200$ 块）：使用分步窗口大小（3，5，8，和 10）分别用于最多 20，50，100，和 200 个块的文档，旨在为较短的上下文保留细节。
- 对于长篇叙述（$N > 200$ ）：应用对数缩放函数以防止窗口过于庞大。此亚线性增长旨在更缓慢地增加对大文本的摘要范围。窗口大小计算如下，以保持窗口大小在 10 到 20 之间：

$$W = \min(20, \max(10, \lfloor \log_2(N) \times 2 \rfloor))$$

对于每个窗口，其中包含的文本块被连接并提供给一个 LLM 代理（我们实验中的 GPT-4o-mini）。代理被指示生成一个简洁的摘要，该摘要保持时间顺序并识别关键事件和因果关系。然后将生成的摘要收集并按原始窗口顺序排序，形成情节层的节点。

B.2 GraphRAG 实验

GraphRAG 是一种类似于 HippoRAGv2 的结构增强 RAG 方法，它涉及从源文档构建一个综合的知识图谱，然后用来识别互联信息以进行检索。然而，其公式化需要大量计算来构建包含多级节点关系和摘要的检索索引。

我们在一个数据子集上进行了初步实验，以评估其可行性。结果如表 5 所示，GraphRAG 不仅消耗的 token 显著更多，且与我们实验中采用的其他基线相比，得分更低。考虑到其计算成本与性能之间的权衡，我们最终未将 GraphRAG 作为全面评估的主要基线。

	ComoRAG	GraphRAG
Performance Metrics		
F1 Score	33.61 (100.0 %)	14.20 (42.3 %)
EM Score	21.43 (100.0 %)	8.00 (37.3 %)
Token Usage		
Tokens	5.90M (100.0 %)	27.12M (459.7 %)
Average Time Taken (sec)		
Index	291 (100.0 %)	1936 (665.3 %)
Retrieve	25 (100.0 %)	29 (116.0 %)

Table 5: ComoRAG 和 GraphRAG 的性能、代币使用情况和平均时间的比较。

B.3 ComoRAG 的超参数

我们的 ComoRAG 框架的关键超参数详见表 6。所有认知代理均采用 GPT-4o-mini，检索由广泛使用的 BGE-M3 嵌入模型提供支持。对于检索设置，动态认知循环被配置为最多运行 5 次迭代，每个循环生成最多 3 个新的探测查询。问答的上下文限制为 6k 个标记，与我们实验中的所有 RAG 基线一致。该上下文通过按比例分配证据进行组装，分别从真实记忆、语义记忆、情景记忆和融合的历史记忆中按 8:2:2:1 的比例进行分配。“Mem-Fuse 阈值”设置为 0.5，表示从记忆池中检索的证据中，有多少比例被转发给整合代理进行记忆融合和摘要生成。

Hyperparameter	Value
LLM Agents (π_{probe} , etc.)	GPT-4o-mini
Retrieval Model	BGE-M3
Chunk Size	512 tokens
Context Length	6,000 tokens
Random Seed	0
Max Iterations	5
Max Probing Queries	3
Context Construction	Proportional Allocation (8:2:2:1 ratio for V:S:E:H)
Mem-Fuse Threshold	0.5

Table 6: 我们实验中 ComoRAG 的超参数设置。V、S、E、H 分别指代真实、语义、情景和历史证据。

C 叙述的查询类型

Dataset	Factoid	Narrative	Inferential	Total
EN.QA	224	84	43	351
EN.MC	132	46	51	229

Table 7: 两组数据中查询类型的分布。

为了便于对我们模型性能进行细粒度分析，我们（本文作者）手动标注了 EN.QA 和 EN.MC 数据集中所有问题的类型。每个问题根据回答问题所需的认知过程被分类为三个类别之一，这些过程在第 4.3 节中描述。

- 事实型：通过从文本中找到单一的、特定的信息来回答的问题。
- 叙事：需要理解情节进展的问题，需要从多个文本部分汇聚信息。
- 推理：需要在字面文本之外进行推理以理解隐含动机或因果联系的问题。

标注的查询类型的最终分布在表 7 中展示。

D 提示模板

Self-Probe Instruction Template for Probing Query Generation in Regulation Agent

Role:
You are an expert in multi-turn retrieval-oriented probe generation. Your job is to extract diverse and complementary retrieval probes from queries to broaden and enrich subsequent corpus search results.

Input Materials:
- 原始查询：需要全面信息检索的问题或信息需求。
- 上下文：可用的背景信息、部分内容或相关摘要。
- 以前的探针：之前迭代中生成的探针（如果有的话）。

Task:
Based on the query and context, generate up to 3 non-overlapping retrieval probes that explore the query from distinct angles.
Critical Requirements:

- 语义区分：确保新探测在语义上与之前提供的任何探测明显不同。
- 互补覆盖：新探针应覆盖先前探针未涉及的不同信息维度。
- 相关性维护：所有探测必须与回答原始查询直接相关。

Each probe should:

- 针对与查询类型相关的不同信息维度：
  - 与角色相关: 行为、动机、关系、时间线、结果
  - 事件相关：参与者、原因、顺序、地点、结果
  - 对象相关：描述、起源、用途、意义、连接
  - 与地点相关：发生的事件、涉及的人、时间段、意义
- Expand search scope beyond obvious keywords to capture related content.
- Avoid semantic overlap with previous probes while maintaining query relevance.
- Be formulated as effective search terms or phrases.

Probe Generation Strategy:

- 当存在之前的探针时：
  1. 分析先前的报道：识别已经覆盖的语义领域/角度。
  2. 差距识别：寻找未探索但相关的信息维度。
  3. 替代角度：从不同的概念视角生成探针。
  4. 语义距离：确保与之前的探测保持足够的语义距离。
- 当不存在先前的探针时：
  - 探针 1：查询中明确提到的直接元素。
  - 探针 2：可能包含答案的上下文元素。
  - 探测 3：相关概念或替代表述。

Output Format:

```json
{
    "probe1": "Content of probe 1",
    ...
}
```

Notes:

- 对于简单的查询，你可能只需生成 1 到 2 个探针。
- 如果以前的探测已经覆盖了大多数相关角度，则生成更少的新探测以避免冗余。
- 优先考虑质量和语义的独特性而非数量。

Mem-Encode Instruction Template for Synthesized Cue Generation in Comprehension Agent

Role
You are an expert narrative analyst capable of identifying, extracting, and analyzing key information from narrative texts to provide accurate and targeted answers to specific questions.

Material
You are given the following:
1. 一个需要解决的最终目标
2. 需要回答的一个具体问题
3. 内容：来自叙述性文本的直接摘录、事实和具体信息

Task
1. 仔细分析问题以识别：
   - 在询问哪种类型的信息（角色动作、地点、对象、事件、动机等）
   - 哪些叙事元素与回答它相关
   - 需要提取的具体细节
2. 系统地扫描内容以查找：
   - 直接提及相关元素（名称、地点、物品、事件）
   - 有助于回答问题的上下文探针
   - 时间和空间关系
   - 因果关系
3. 分析所识别的信息，考虑：
   - 明确陈述（直接陈述的事实）
   - 隐含信息（通过上下文、对话或叙述暗示）
   - 不同叙述元素之间的逻辑联系
   - 如果相关的话，事件的时间顺序
4. 综合研究结果以构建对问题的精确答案。

Response Format
Provide a structured analysis with up to 5 key findings:
‘‘‘
Key Finding : <Most directly relevant information answering the question>
Key Finding : <Supporting evidence or context>
Key Finding : <Additional relevant details>
Key Finding : <Clarifying information if needed>
Key Finding : <Resolution of any ambiguities>
‘‘‘

Mem-Fuse Instruction Template for Cue Generation in Integration Agent

Role:
You are an expert narrative synthesis specialist who excels at integrating and analyzing information from multiple narrative sources to create coherent and comprehensive insights.

Input Material:
- 先前分析：来自早期记忆融合操作的结果，其中包含已分析的叙述信息。
- 当前查询：需要解决的问题或信息请求。

Task:
1. 回顾并理解之前的记忆融合输出：
   - 识别关键叙述元素及其关系。

- 注意任何已确立的事实、人物发展或情节要点。
- 识别不同分析中的模式和联系。

2. 分析当前查询的上下文：
   - 确定其与先前建立的信息的关系。
   - 确定需要解决的任何新方面或角度。
   - 考虑以前的见解如何能够为当前的回应提供信息。
3. 综合信息：
   - 将相关的以往研究结果与新的分析相结合。
   - 创建一个连贯的叙述来回答当前的查询。
   - 确保与先前分析的连续性和一致性。
   - 突出显示任何新的见解或发展。
4. 提供一个全面的回答：

Response Format:
Provide a cohesive narrative response that integrates previous insights with new analysis to address the current query. Focus on creating a flowing, well-structured response.

Try-Answer Prompt Template for Query Resolution in QA Agent

Role:
You are an expert on reading and understanding books and articles.

Task:
Given the following detailed article, semantic summary, Episodic summary from a book, and a related question with different options, you need to analyze which option is the best answer for the question.

Inputs:
- 详细文章： { context }
- 语义总结： { semantic_summary }
- 按照情节总结： { Episodic_summary }
- 历史信息： { history_info }
- 问题： { question }

Limits:
- 不要推断。只严格基于提供的内容作出回应。
- 只有在至少找到两个地方支持答案的情况下才选择这一项。

Response Format:
1. 内容理解：
   以不超过三句话的简要总结内容开始。请以 # # # Content Understanding 开始这一部分。
2. 问题分析：
   根据问题，使用 markdown 列表分析并列出所有相关项。本节以 # # # Question Analyse 开始。
3. 期权分析：
   提取与 4 个选项相关的关键点，同时使用 markdown 列表。以 # # # Options Analyse 开始此部分。
   注意：仅基于提供的材料进行分析，不要进行猜测。
4. 最终答案：
   请在标题中提供您的最终答案。这个部分以 # # # Final Answer 开始，随后是以 [A] 或 [B] 或 [C] 或 [D] 格式的最佳选项。如果您无法回答，给出一个失败信号： * 。

E 叙述推理个案研究

Input Data (No Options)

Query: Mrs. MacIntyre never writes letters, so what is the sudden purpose of buying ink?
Options: [A] Response to the government's letter [B] Sending a birthday card to my niece [C] Write a letter to the Sunday Comet newspaper. [D] Blurring Photos

ComoRAG's Choice Result

Memory Pool $M_{pool}^{(0)}$:
- A page of the newspaper had a portion cut out...

Step1

Probes $P^{(1)}$:
- What did Mrs. McGinty recognize in the clipping, and how did she try to profit from it?
...
Retrieved Passages:
...The narrative offers insight into Miss Pamela Hosford' s role at the Sunday Comet, as she casually recalls receiving a vague letter from a woman about a photograph but fails to retrieve it...
Cues $C^{(1)}$:
- Key Finding:Mrs. McGinty usually had Joe help her reply to letters.;
- Key Finding:Mrs. McGinty may have sent a letter to the newspaper.;...

Memory Pool $M_{pool}^{(1)}$:
- A page of the newspaper had a portion cut out...
- Mrs. MacIntyre sent a letter to the Sunday Comet...

Step2

Probes $P^{(2)}$:
- What did Miss Pamela Hosford recall about a letter sent to the Sunday Comet, and what might it imply about Mrs. McGinty?
...
Retrieved assages:
...Miss Pamela Hosford' s role at the Sunday Comet, as she casually recalls receiving a vague letter from a woman about a photograph but fails to retrieve it...She liked eavesdropping, knew who was in the photo, and asked what we ' pay...
Cues $C^{(2)}$:
- Key Finding:Mrs. McGinty wanted to make some money from the photo.;...
Chosen: 正确
(C) Write a letter to the Sunday Comet newspaper: Strong textual probes support this option. Mrs. McGinty cut out a part of the newspaper, recognized someone in a photo, asked about payment, and unusually bought ink—suggesting she intended to write to the paper. 最终答案：[C]

Table 8: 叙事推理案例研究。我们展示了一个案例来证明我们模型在长文本理解方面的性能，展示了元认知控制环路的最后一轮。不同的颜色用于突出处理信息的性质： Blue 用于表明有助于正确答案的关键证据，而 Orange 用于表明关键线索。