Metacognitive Multi-Agent Retrieval-Augmented Generation for Multi-Hop Reasoning

Gabriel Iturra-Bocaz

University of Stavanger, Norway
gabriel.e.iturrabocaz@uis.no

Abstract. Retrieval-Augmented Generation (RAG) enhances Large Language Models by grounding their responses in external knowledge, yet existing approaches struggle with complex reasoning tasks that require combining multiple sources of information. Although recent multi-retrieval methods integrate iterative retrieval and reasoning, they still lack mechanisms to decide when to stop retrieving and how to maintain reasoning coherence, often leading to error propagation and hallucinations. This research introduces a metacognitive multi-agent framework for RAG that models reasoning as a collaborative process guided by metacognitive control and shared memory systems. The framework enables dynamic coordination between retrieval and reasoning, allowing the system to monitor its progress, assess evidence sufficiency, and revise its reasoning when inconsistencies appear. By incorporating metacognitive regulation and explicit memory interaction, the proposed approach aims to improve reasoning reliability, factual grounding, and interpretability in multi-hop question answering.

Keywords: Language Agents · Retrieval-Augmented Generation · Multi Hop Questions.

1 Introduction

Retrieval-Augmented Generation (RAG) [10,4] is a key approach for improving the factual accuracy of Large Language Models (LLMs) [28] by using external information during generation. Recent work has introduced multi-retrieval methods [22,29,19], where retrieval and reasoning interact over several steps to solve complex tasks [5], including multi-hop questions [11]. However, these systems often lack awareness of when to stop retrieving information [24], continuing to gather evidence without recognizing when it is sufficient to answer the question. Because retrieval and reasoning are closely linked [5], errors in one can reinforce errors in the other, leading to hallucinations [7] and reducing the overall reliability of responses.

Human reasoning is self-aware and adaptable, a capacity known as metacognition [8,18], which helps people improve their thinking, planning, and learning. Cognitive Architectures [27] such as Soar [9], ACT-R [16], and Cognitive Architectures for Language Agents (CoAla) [21] have extensively examined these processes. They suggest that intelligence arises from the coordination of specialized components distributed across multiple memory systems, including working, semantic, and episodic memory, all regulated by a central control mechanism that monitors, evaluates, and adjusts cognitive activity. These metacognitive abilities enable humans to plan ahead, detect uncertainty, and revise their strategies dynamically, ensuring that reasoning remains efficient and grounded.

Inspired by these principles, this research proposes a metacognitive multi-agent framework for RAG that follows a human-like coordination process [2,17]. The framework includes four specialized agents, the Planner, Retriever, Reasoner, and Verifier, which work together through shared memory systems, including working, semantic, and episodic memory, all supervised by a metacognitive controller. Each agent has a specific role, while the controller keeps the process coherent, decides when more evidence is needed, and fixes inconsistencies. Unlike existing RAG systems, this approach introduces explicit metacognitive control and memory-based coordination, enabling the model to regulate its own retrieval and reasoning process for more reliable and interpretable responses.

2 Related Work

Early RAG systems relied on a single retrieval pass, which limited their ability to reason across multiple, interdependent facts. To address this, several studies introduced iterative and multi-retrieval strategies that alternate between reasoning and retrieval. Yao et al. [26] proposed ReAct, a framework that integrates reasoning and acting by allowing models to produce both reasoning traces and retrieval actions in the same loop. Trivedi et al. [22] extended this idea through interleaved chain-of-thought reasoning and retrieval, where intermediate reasoning steps guide evidence selection. Press et al. [14] introduced Self-Ask, which decomposes complex questions into sub-queries answered iteratively with retrieval support. Although these methods improve retrieval quality and reasoning depth, they still lack self-regulation, leading to redundant searches and unstructured reasoning in complex multi-hop tasks.

Parallel to multi-retrieval RAG systems, there is growing interest in multi-agent and reflective reasoning systems. Frameworks such as Reflexion [19] and Self-RAG [1] introduce feedback loops that allow models to critique and refine their own outputs. Multi-agent approaches [13,12,17] further distribute reasoning across specialized roles that cooperate to solve complex tasks. However, these systems often lack a unified control mechanism to coordinate agent interactions and regulate the balance between retrieval and reasoning, leaving communication largely heuristic and prone to over-retrieval or inconsistency.

Research in Cognitive Architectures provides a foundation for designing systems capable of adaptive reasoning and self-regulation. Classical architectures such as ACT-R [16] and Soar [9] implement cognition through symbolic, rule-based agents that coordinate perception, memory, and action under an executive control system. More recent architectures, such as CoALA [21] and Cognitive LLMs [23], extend these ideas to language-based systems, using large language models as cognitive components that communicate and cooperate through natural language rather than production rules. While these architectures focus on cognitive control and modular reasoning, they do not explicitly incorporate metacognitive self-evaluation or reasoning oversight. The proposed research builds upon this gap by introducing metacognitive regulation within a multi-agent RAG framework, enabling the system to monitor reasoning progress, assess evidence sufficiency, and dynamically adjust retrieval or inference strategies.

3 Research Questions and Proposed Methodology

This research aims to explore how metacognitive control and cognitive architecture principles can be extended to RAG through a multi-agent system capable of adaptive reasoning and self-regulation. The central objective is to design a framework that coordinates multiple specialized agents-each responsible for planning, retrieval, reasoning, and verification-under the supervision of a metacognitive controller. To achieve this, we address the following research questions:

3.1 RQ1: How can multiple agents be coordinated to solve complex tasks such as multi-hop questions?

We hypothesize that principles from Cognitive Architectures [9,16,21,23] can guide the design of collaborative, metacognitively regulated multi-agent systems. Classical architectures such as Soar [9] and ACT-R [16] organize cognition around modular, specialized components (e.g., working memory, procedural memory) that are coordinated by a central executive process [21]. Inspired by these models, we propose a **metacognitive multi-agent RAG** framework where agents operate as distributed cognitive modules connected through shared memory systems.

The system will include four five agents—*Planner*, *Retriever*, *Reasoner*, *Verifier* and, *Generator*—and a *Metacognitive Controller*. The controller monitors agent interactions, evaluates reasoning progress, and determines when retrieval or reasoning should continue or terminate. This architecture emphasizes coordination and control rather than internal cognitive modeling of each agent, bridging cognitive architectures with cooperative agent systems.

3.2 RQ2: How can agent-based RAGs integrate retrieval to better support reasoning and planning?

In traditional RAG pipelines [10,4], retrieval operates passively-queries are generated from the initial question or reasoning steps, but the system lacks awareness of when information is sufficient [24]. Our proposed framework introduces retrieval as a dynamic, goal-directed process, coordinated by the Planner and monitored by the Metacognitive Controller.

At each reasoning step, the Planner generates hypotheses or subgoals, which guide the Retriever to issue targeted queries. Retrieved evidence is then stored in a working memory buffer, accessible to the Reasoner. The Controller evaluates whether the retrieved evidence sufficiently supports ongoing reasoning, using explicit metacognitive signals such as uncertainty, redundancy, or contradiction. This enables retrieval to be adaptive—triggered only when necessary—and ensures that reasoning remains coherent and grounded rather than over-retrieving or drifting off-topic.

3.3 RQ3: How can the reasoning quality and factual grounding of agent-driven RAG systems be evaluated?

To evaluate the proposed framework, we will combine standard QA metrics with metrics that measure how retrieved documents contribute to individual reasoning steps. Traditional metrics such as Exact Match (EM) [15] and Answer Accuracy will assess final output quality, while additional metrics will capture how effectively retrieved evidence supports reasoning steps.

In particular, we plan to adopt and extend existing metrics such as:

- **Faithfulness** [3], measuring the degree to which generated answers are grounded in retrieved evidence;
- **Context Precision/Recall** [3], evaluating whether retrieved passages contain information explicitly used in the reasoning chain.

We also plan to design and evaluate new retrieval-reasoning metrics that explicitly quantify how retrieval contributes to each reasoning step in multi-hop question answering tasks, and how the reasoning steps generated by planner LLMs can be leveraged by executor LLMs. These metrics will extend current analyses from open-domain benchmarks such as HotpotQA [25], 2WikiMultiHopQA [6] and BRIGHT [20], as well as internal datasets from our ongoing projects, including those focused on clinical reasoning and temporal questions. The goal is to capture the interaction between evidence selection and reasoning quality, enabling an evaluation framework that measures not only answer correctness but also the factual grounding and reasoning sufficiency of multi-agent RAG systems.

4 Conclusion

This research proposes a novel integration of cognitive and metacognitive principles into multi-agent RAG systems. By modeling reasoning as a collaborative process guided by shared memories and a central metacognitive controller, the framework seeks to achieve adaptive retrieval, consistent reasoning, and reliable self-correction. The expected outcome is a system capable of regulating its own reasoning process—improving factual grounding, interpretability, and robustness in complex multi-hop tasks—thereby advancing the current frontier of reasoning-aware and self-reflective language models.

Disclosure of Interests

The authors declare that they have no competing interests relevant to the content of this article.

References

1. Asai, A., Wu, Z., Wang, Y., Sil, A., Hajishirzi, H.: Self-rag: Learning to retrieve, generate, and critique through self-reflection (2024)
2. Du, Y., Leibo, J.Z., Islam, U., Willis, R., Sunehag, P.: A review of cooperation in multi-agent learning. arXiv preprint arXiv:2312.05162 (2023)
3. Es, S., James, J., Anke, L.E., Schockaert, S.: Ragas: Automated evaluation of retrieval augmented generation. In: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations. pp. 150–158 (2024)
4. Fan, W., Ding, Y., Ning, L., Wang, S., Li, H., Yin, D., Chua, T.S., Li, Q.: A survey on rag meeting llms: Towards retrieval-augmented large language models. In: Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining. pp. 6491–6501 (2024)
5. Gao, Y., Xiong, Y., Zhong, Y., Bi, Y., Xue, M., Wang, H.: Synergizing rag and reasoning: A systematic review. arXiv preprint arXiv:2504.15909 (2025)
6. Ho, X., Nguyen, A.K.D., Sugawara, S., Aizawa, A.: Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. arXiv preprint arXiv:2011.01060 (2020)
7. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al.: A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems **43**(2), 1–55 (2025)
8. Lai, E.R.: Metacognition: A literature review (2011)
9. Laird, J.E., Newell, A., Rosenbloom, P.S.: Soar: An architecture for general intelligence. Artificial intelligence **33**(1), 1–64 (1987)
10. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems **33**, 9459–9474 (2020)
11. Mavi, V., Jangra, A., Jatowt, A., et al.: Multi-hop question answering. Foundations and Trends® in Information Retrieval **17**(5), 457–586 (2024)
12. Nguyen, T., Chin, P., Tai, Y.W.: Ma-rag: Multi-agent retrieval-augmented generation via collaborative chain-of-thought reasoning. arXiv preprint arXiv:2505.20096 (2025)
13. Park, J.S., O'Brien, J., Cai, C.J., Morris, M.R., Liang, P., Bernstein, M.S.: Generative agents: Interactive simulacra of human behavior. In: Proceedings of the 36th annual acm symposium on user interface software and technology. pp. 1–22 (2023)
14. Press, O., Zhang, M., Min, S., Schmidt, L., Smith, N.A., Lewis, M.: Measuring and narrowing the compositionality gap in language models. In: Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 5687–5711 (2023)
15. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250 (2016)

16. Ritter, F.E., Tehranchi, F., Oury, J.D.: Act-r: A cognitive architecture for modeling cognition. Wiley Interdisciplinary Reviews: Cognitive Science **10**(3), e1488 (2019)
17. Salemi, A., Maddipatla, M., Zamani, H.: Ciir@ liverag 2025: Optimizing multi-agent retrieval augmented generation through self-training. arXiv preprint arXiv:2506.10844 (2025)
18. Schraw, G., Moshman, D.: Metacognitive theories. Educational psychology review **7**(4), 351–371 (1995)
19. Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., Yao, S.: Reflexion: Language agents with verbal reinforcement learning. Advances in Neural Information Processing Systems **36**, 8634–8652 (2023)
20. Su, H., Yen, H., Xia, M., Shi, W., Muennighoff, N., Wang, H.y., Liu, H., Shi, Q., Siegel, Z.S., Tang, M., et al.: Bright: A realistic and challenging benchmark for reasoning-intensive retrieval. arXiv preprint arXiv:2407.12883 (2024)
21. Sumers, T., Yao, S., Narasimhan, K., Griffiths, T.: Cognitive architectures for language agents. Transactions on Machine Learning Research (2023)
22. Trivedi, H., Balasubramanian, N., Khot, T., Sabharwal, A.: Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In: Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers). pp. 10014–10037 (2023)
23. Wu, S., Oltramari, A., Francis, J., Giles, C.L., Ritter, F.E.: Cognitive llms: Toward human-like artificial intelligence by integrating cognitive architectures and large language models for manufacturing decision-making. Neurosymbolic Artificial Intelligence **1**, 29498732251377341 (2025)
24. Yang, D., Zeng, L., Rao, J., Zhang, Y.: Knowing you don't know: Learning when to continue search in multi-round rag through self-practicing. In: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1305–1315 (2025)
25. Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W., Salakhutdinov, R., Manning, C.D.: Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In: Proceedings of the 2018 conference on empirical methods in natural language processing. pp. 2369–2380 (2018)
26. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K.R., Cao, Y.: React: Synergizing reasoning and acting in language models. In: The eleventh international conference on learning representations (2022)
27. Ye, P., Wang, T., Wang, F.Y.: A survey of cognitive architectures in the past 20 years. IEEE transactions on cybernetics **48**(12), 3280–3290 (2018)
28. Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al.: A survey of large language models. arXiv preprint arXiv:2303.18223 **1**(2) (2023)
29. Zhou, Y., Liu, Z., Jin, J., Nie, J.Y., Dou, Z.: Metacognitive retrieval-augmented large language models. In: Proceedings of the ACM Web Conference 2024. pp. 1453–1463 (2024)