Language Agents for Information Retrieval

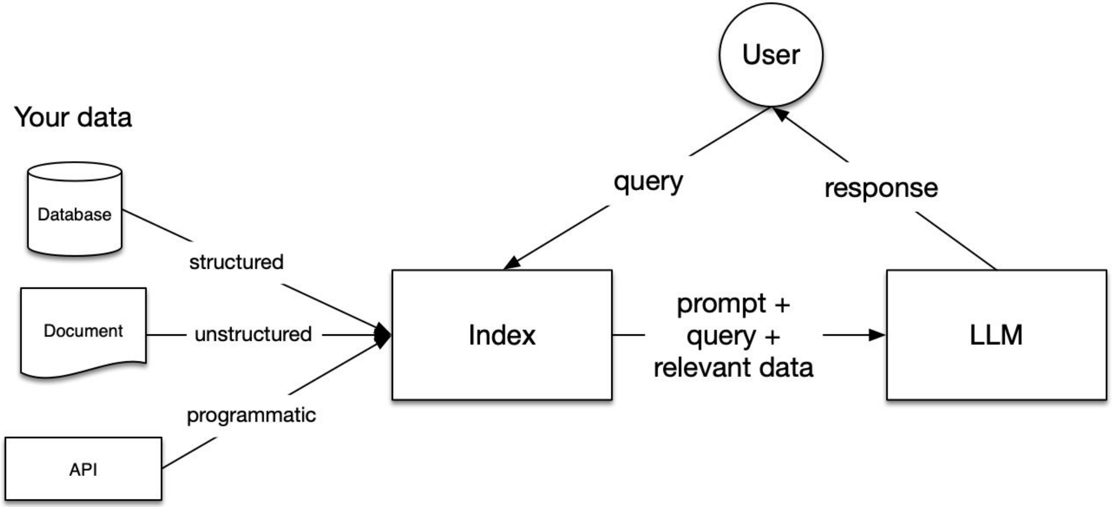
Gabriel Iturra-Bocaz

gabriel.e.iturrabocaz@uis.no

https://giturra.cl/

University of Stavanger

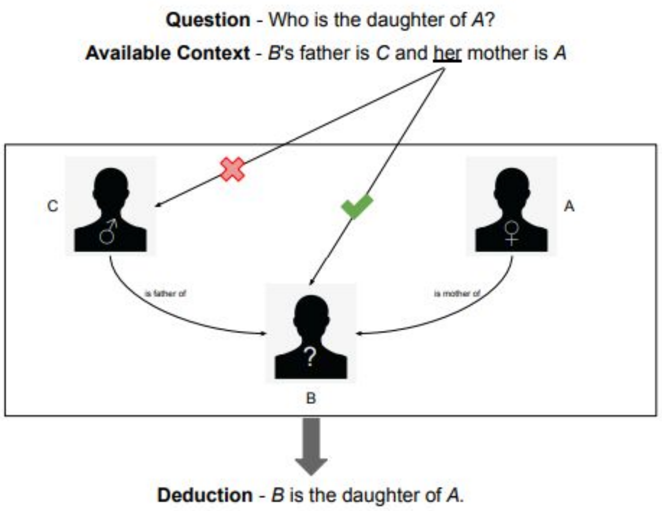
RAG Challenges



- A **retrieval model** recalls relevant documents from an external knowledge source.
- A **language model** then uses the recalled context to generate informed, grounded responses.
- Performs well on factual questions.

Challenges of RAG

- **Limited reasoning:** focuses on single-query retrieval with no global or multi-step logic.
- **Fragmented context:** recalled passages are partial or disconnected.
- **No continuity:** lacks planning and memory, struggling with multi-hop questions.



Question - Who is the daughter of *A*?

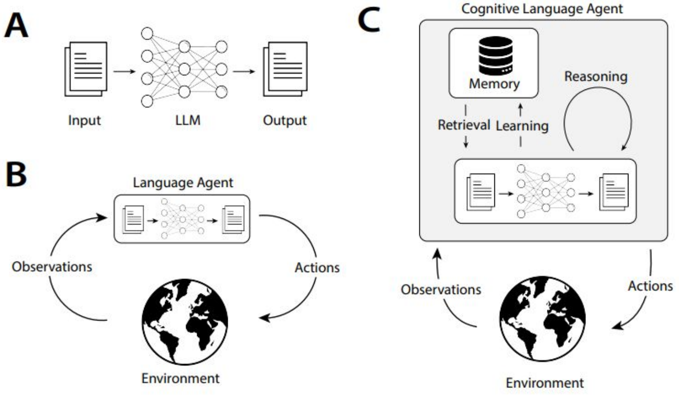
Available Context - *B*'s father is *C* and her mother is *A*

Deduction - *B* is the daughter of *A*.

https://www.nowpublishers.com/article/Details/INR-102

Language Agents

An autonomous agent is a system situated within and apart of an environment that senses that environment and acts on it, over time, in pursuit of its own agenda and so as to effect what it senses in the future.



https://arxiv.org/abs/2309.02427

From RAG to Agent-Based Systems

- The research community is moving toward agent-based RAGs.
- Agents **can plan, reason, and coordinate** to handle complex queries.
- **Agent-RAG frameworks** integrate retrieval with reasoning and reflection.

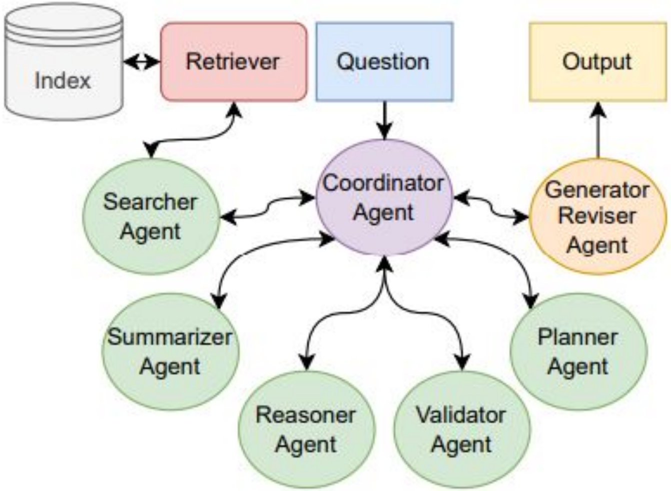


Figure 1: Overview of multi-agent RAG.

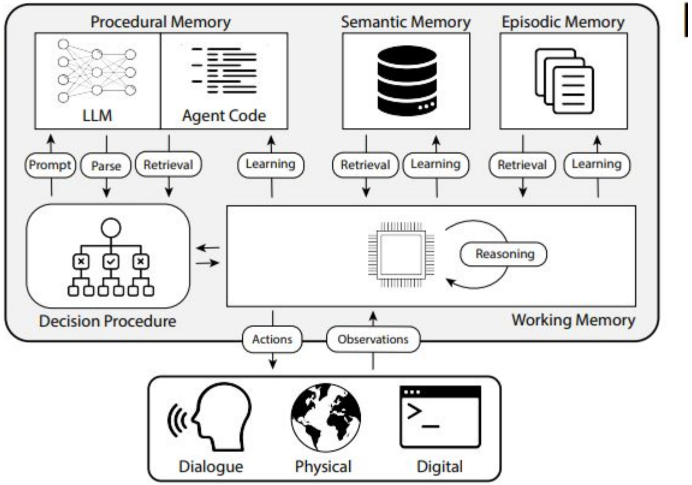
https://arxiv.org/pdf/2506.10844

Research Questions

1. How can multiple agents be coordinated to solve complex tasks like multi-hop questions?
2. How can agent-based RAGs integrate retrieval to better support reasoning and planning?
3. How can the reasoning quality and factual grounding of agent-driven RAG systems be evaluated?

Cognitive Architectures

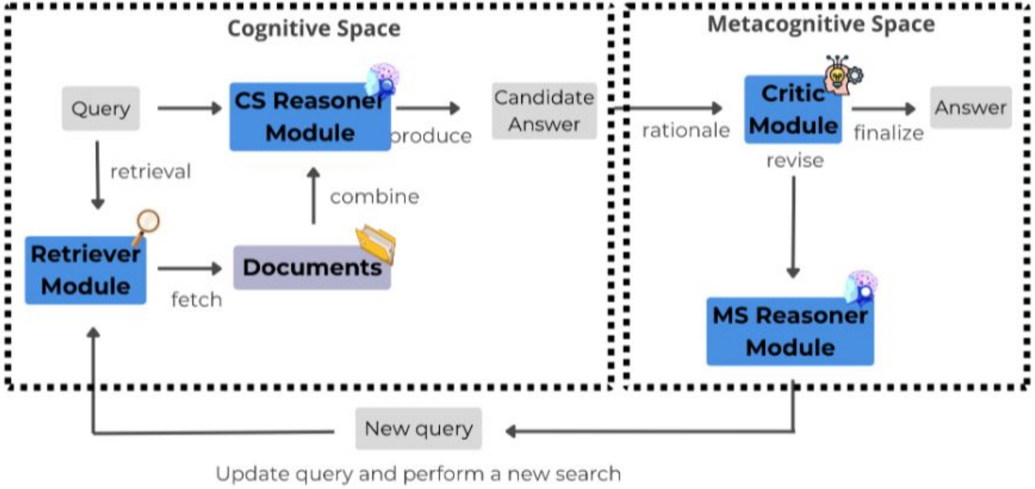
- To address **RQ1**, we hypothesize that organizing agents using principles of cognitive architectures can enhance retrieval integration.
- Agents take on specialized roles (retrieval, reasoning, planning) and interact through a shared memory or control mechanism.
- This structure aims to enable more coherent reasoning and decision-making across complex tasks.



https://arxiv.org/abs/2309.02427

Towards Metacognition-Aware Retrieval-Augmented Generation

- To answer **RQ2**, we apply metacognitive principles to RAG, enabling adaptive reasoning and self-regulated retrieval.
- Metacognition aims RAG systems avoid redundant retrieval, reduce hallucinations, and balance efficiency with reasoning accuracy.



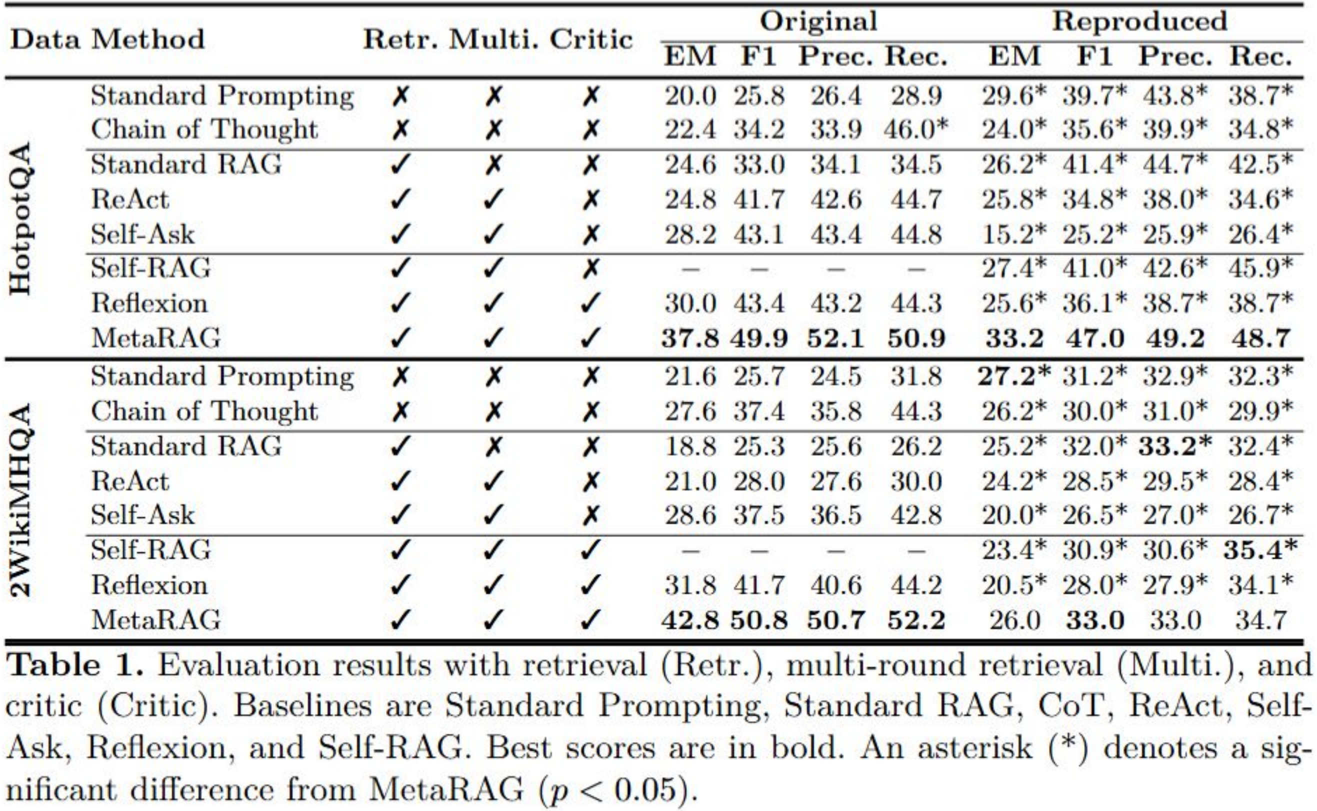
Towards Metacognition-Aware Retrieval-Augmented Generation

- We started exploring metacognition in RAG systems by reproducing and comparing **MetaRAG (WWW 2024)** and **SIM-RAG (SIGIR 2025)**.
- MetaRAG proposes an explicit monitor–evaluate–plan loop, enabling structured self-evaluation and adaptive reasoning.
- SIM-RAG focuses on implicit self-reflection through iterative retrieval and refinement.
- By comparing both approaches, we aim to understand how different forms of metacognitive control impact retrieval efficiency and reasoning quality in RAG systems.

A Reproducibility Study of Metacognitive Retrieval-Augmented Generation

- We reimplemented MetaRAG under its original experimental setup, reproducing multi-hop QA experiments on HotpotQA and 2WikiMultiHopQA.
- We employed **GPT-3.5-turbo-16k** and **Llama-3.3-70B** as the reasoning backbones in both cognitive and metacognitive spaces.
- We conducted a reproducibility evaluation to assess alignment with the original results.
- By comparing both approaches, we aim to understand how different forms of metacognitive control impact retrieval efficiency and reasoning quality in RAG systems.
- Our findings show that MetaRAG's relative improvements over RAG and reasoning baselines remain consistent.
- This work has been submitted to ECIR 2026.

A Reproducibility Study of Metacognitive Retrieval-Augmented Generation



Data	Method	Retr.	Multi.	Critic	Original				Reproduced			
					EM	F1	Prec.	Rec.	EM	F1	Prec.	Rec.
HotpotQA	Standard Prompting	✗	✗	✗	20.0	25.8	26.4	28.9	29.6*	39.7*	43.8*	38.7*
	Chain of Thought	✗	✗	✗	22.4	34.2	33.9	46.0*	24.0*	35.6*	39.9*	34.8*
	Standard RAG	✓	✗	✗	24.6	33.0	34.1	34.5	26.2*	41.4*	44.7*	42.5*
	ReAct	✓	✓	✗	24.8	41.7	42.6	44.7	25.8*	34.8*	38.0*	34.6*
	Self-Ask	✓	✓	✗	28.2	43.1	43.4	44.8	15.2*	25.2*	25.9*	26.4*
	Self-RAG	✓	✓	✗	–	–	–	–	27.4*	41.0*	42.6*	45.9*
	Reflexion	✓	✓	✓	30.0	43.4	43.2	44.3	25.6*	36.1*	38.7*	38.7*
	MetaRAG	✓	✓	✓	**37.8**	**49.9**	**52.1**	**50.9**	**33.2**	**47.0**	**49.2**	**48.7**
2WikiMHQA	Standard Prompting	✗	✗	✗	21.6	25.7	24.5	31.8	**27.2***	31.2*	32.9*	32.3*
	Chain of Thought	✗	✗	✗	27.6	37.4	35.8	44.3	26.2*	30.0*	31.0*	29.9*
	Standard RAG	✓	✗	✗	18.8	25.3	25.6	26.2	25.2*	32.0*	**33.2***	32.4*
	ReAct	✓	✓	✗	21.0	28.0	27.6	30.0	24.2*	28.5*	29.5*	28.4*
	Self-Ask	✓	✓	✗	28.6	37.5	36.5	42.8	20.0*	26.5*	27.0*	26.7*
	Self-RAG	✓	✓	✓	–	–	–	–	23.4*	30.9*	30.6*	**35.4***
	Reflexion	✓	✓	✓	31.8	41.7	40.6	44.2	20.5*	28.0*	27.9*	34.1*
	MetaRAG	✓	✓	✓	**42.8**	**50.8**	**50.7**	**52.2**	26.0	**33.0**	33.0	34.7

Table 1. Evaluation results with retrieval (Retr.), multi-round retrieval (Multi.), and critic (Critic). Baselines are Standard Prompting, Standard RAG, CoT, ReAct, Self-Ask, Reflexion, and Self-RAG. Best scores are in bold. An asterisk (*) denotes a significant difference from MetaRAG ($p < 0.05$).

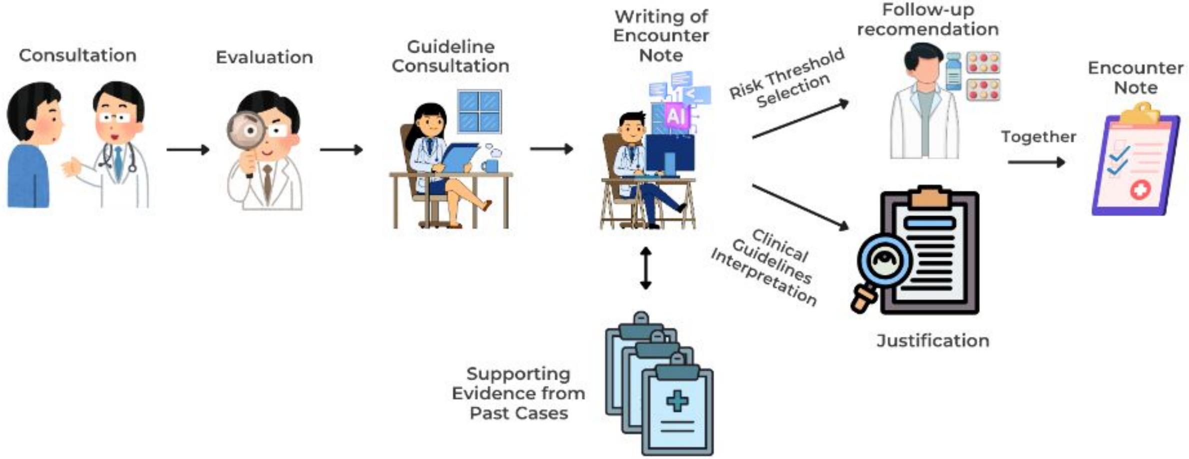
A Reproducibility Study of Metacognitive Retrieval-Augmented Generation

- Reranker Effects
  - Integrating PointWise (BGE) and ListWise (RankGPT) rerankers improved MetaRAG across both datasets.
  - Rerankers reduced retrieval noise and enhanced reasoning accuracy.
  - Llama-3.3-70B consistently outperformed GPT-3.5-turbo-16k.

- MetaRAG vs. SIM-RAG
  - MetaRAG outperformed SIM-RAG under all retrieval and reranking settings.
  - SIM-RAG's fine-tuned critic degraded when rerankers changed retrieval order.
  - MetaRAG's prompt-based control proved more robust and adaptable across models.

From Structured Data to Question-Answer Datasets in Early Clinical Decision-Making

- To answer **RQ3**, we propose creating a multi-hop QA dataset in the medical domain.
- We design the dataset to replicate early-stage clinical reasoning by GPs before a formal diagnosis is made.
- This work has been submitted to LREC 2026.



From Structured Data to Question-Answer Datasets in Early Clinical Decision-Making

- In creating the dataset, we use LLMs to generate synthetic queries and answers that reflect how GPs reason over EHR data.
- We used EHR data from the PI-CAI dataset as the foundation for our study.

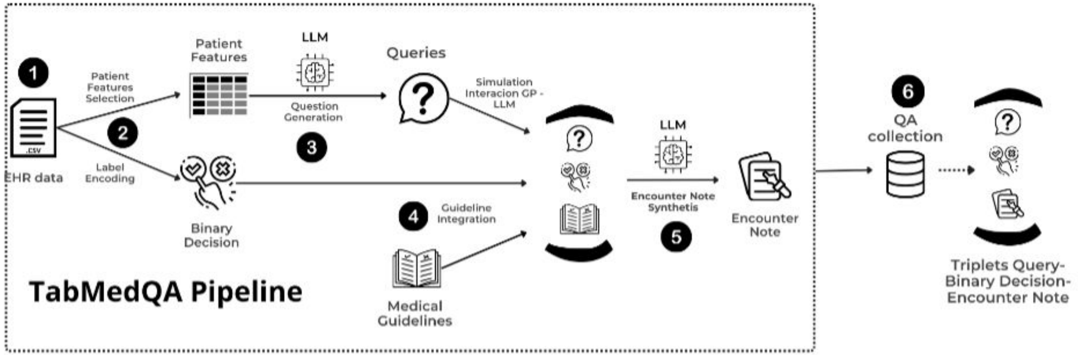


Figure 2: Overview of the **TabMedQA** pipeline: From structured (1) EHR data, (2) *EHR Preprocessing* (patient feature selection and label encoding), (3) *Question Generation*, (4–5) *Guideline Integration*, *Encounter Note Synthesis*, and (6) the resulting *QA collection*.

Future steps

- Create a dataset for multi-hop temporal questions.
- Propose a new cognitive architecture for reasoning in temporal and multi-hop settings.
- Introduce a complementary reasoning framework based on RAG agents, aligned with the proposed cognitive architecture.
- Evaluate our methodology and reasoning capabilities on the newly created dataset.

Questions?

Language Agents for Information Retrieval

Gabriel Iturra-Bocaz
gabriel.e.iturrabocaz@uis.no
https://giturra.cl/
University of Stavanger