Retrieval-Augmented Generation with Conflicting Evidence

Han Wang Archiki Prasad Elias Stengel-Eskin Mohit Bansal
University of North Carolina at Chapel Hill
{hwang, archiki, esteng, mbansal}@cs.unc.edu

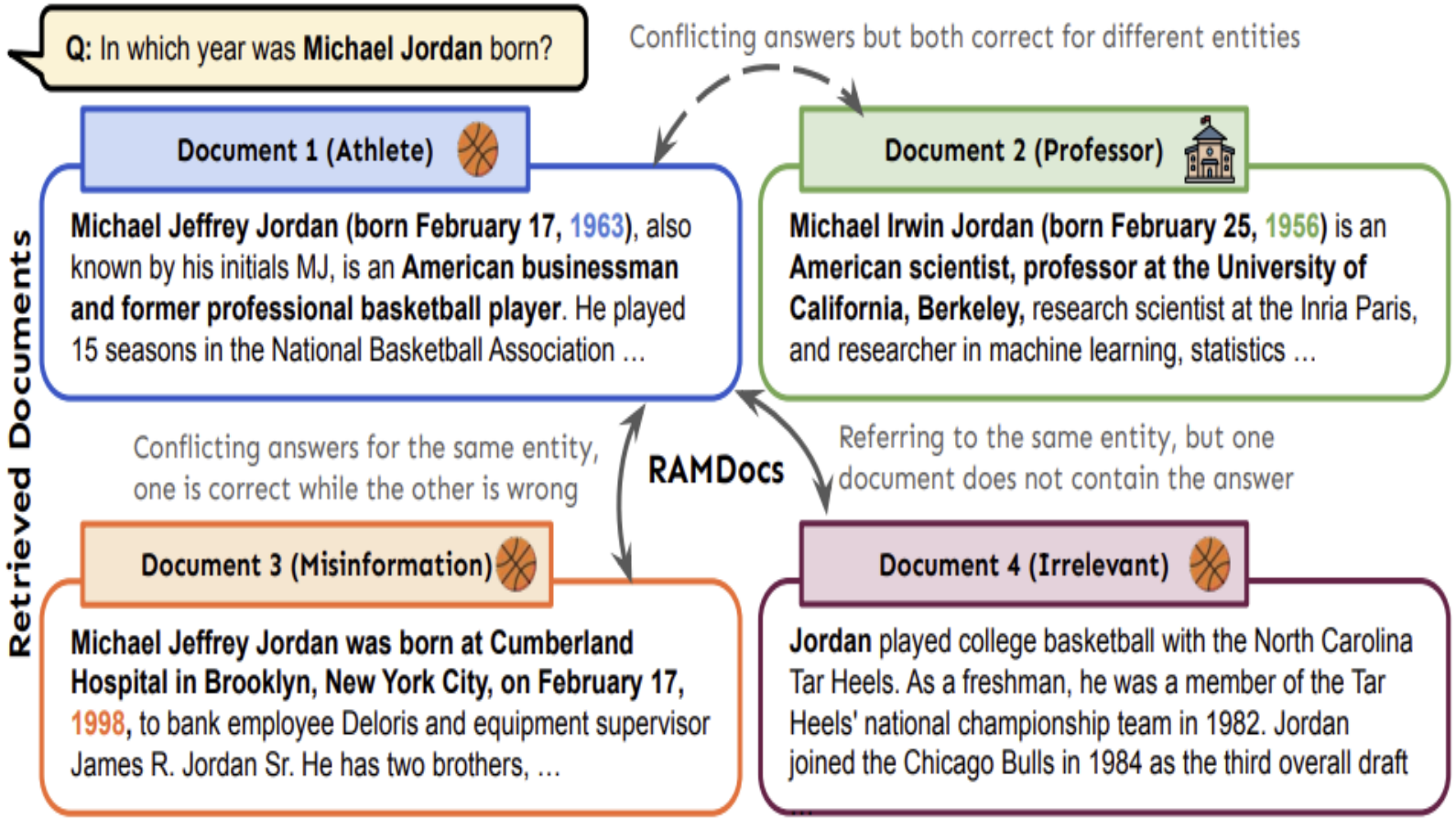
고경빈

2025.05.02

Background

- RAG enables LLMs to generate more accurate and reliable responses
- Online info can be unreliable and unclear query brings in conflicting answers
- Existing researches/benchmarks show only one type of conflict
- Existing methods try to remove bad content to improve RAG
- Choosing one answer fails if many are correct

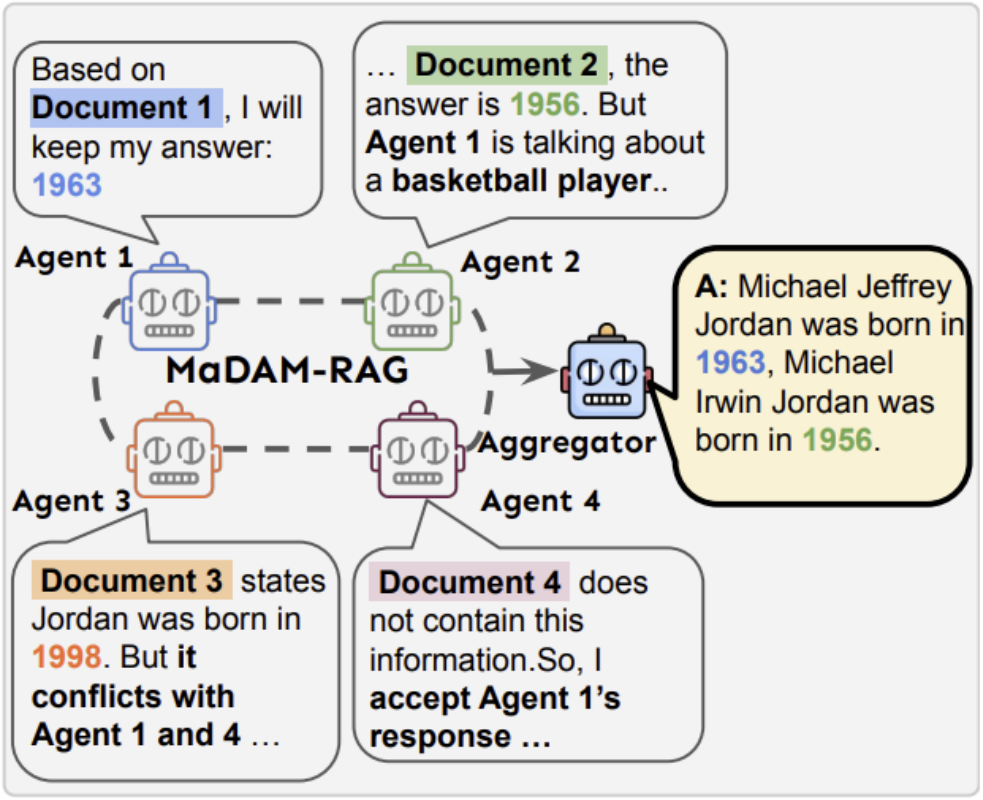
RAMDocs: Retrieval with Ambiguity & Misinformation in Documents



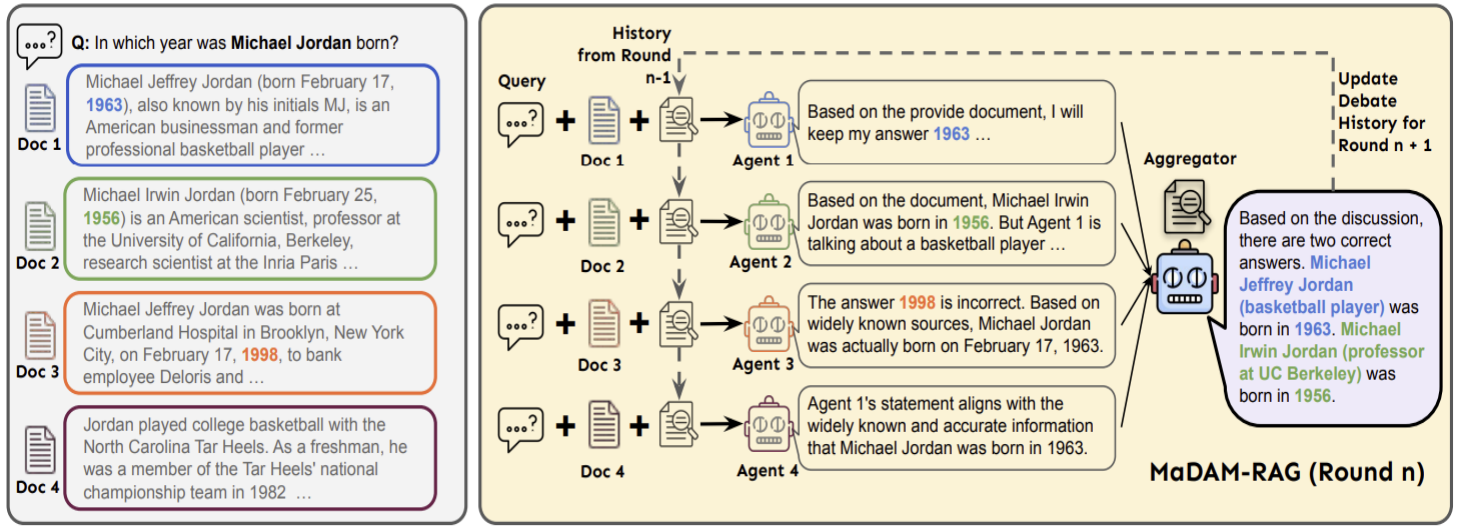
RAMDocs

- Ambiguous Queries
  - Randomly sample 1 to 3 correct answers per ambiguous query, making 500 queries total
- Distribution of Supporting Documents
  - Find multiple documents per query, keep only chunks with the correct answer, randomly assign 1–3 supporting documents
- Misinformation and Noisy Documents
  - Add documents with misinformation and noise to each query

MADAM-RAG(Multi-agent Debate for Ambiguity and Misinformation in RAG)



MADAM-RAG



- Independent LLM dialogue agents: $r_i = \mathcal{L}(q, d_i) \rightarrow \mathcal{R}^{(t)} = \left\{ r_i^{(t)} \right\}_{i=1}^{n}$
- Aggregator: $\left(y^{(t)}, e^{(t)} \right) = \mathcal{A}(\mathcal{R}^{(t)})$
- Iterative multi-round debate process: $r_i^{(t)} = \mathcal{L}_i(q, d_i, y^{(t-1)}, e^{(t-1)})$
- Early stopping: $\forall i, r_i^{(t)} = r_i^{(t-1)}$

Experimental Setup

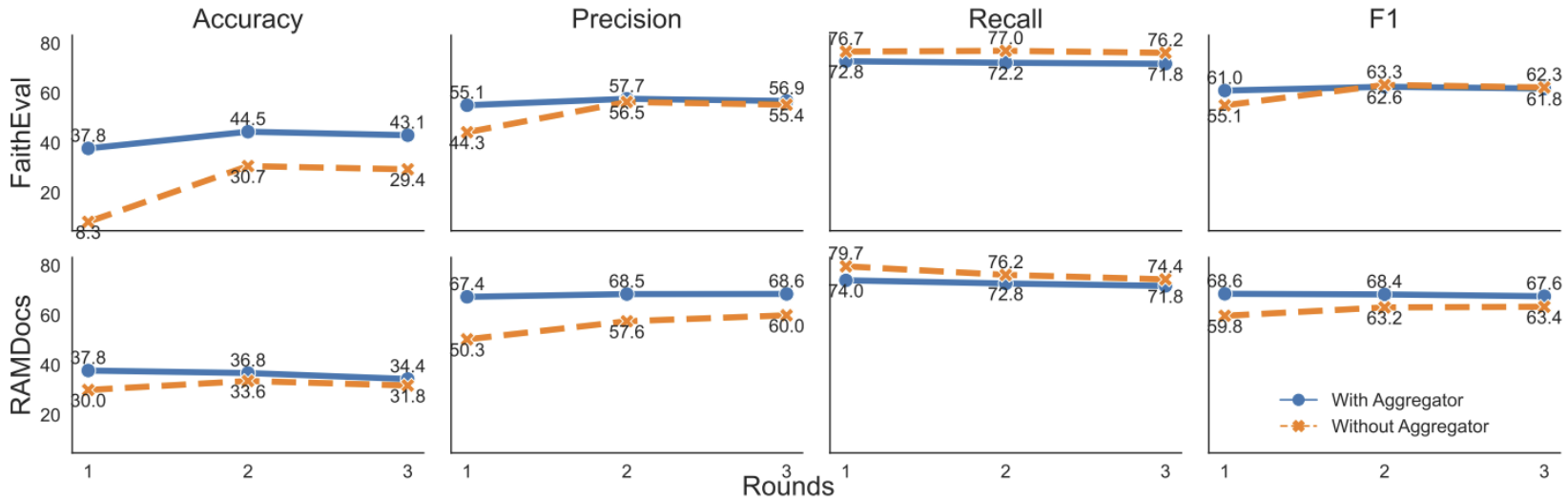
- Datasets
  - FaithEval: test if LLMs stay accurate when evidence includes misinformation(1000)
  - AmbigDocs: check if questions have documents with different correct answers(1000)
  - RAMDocs: test real-world cases with multiple answers, conflicts, and noise(500 from AmbigDocs)
- Metrics: Exact Match
- Models
  - Llama3.3-70B-Instruct, Qwen2.5-72B-Instruct
  - GPT-4o-mini
- Baselines(T=3)
  - No RAG: prompt the LLM with the question only
  - Concatenated-prompt: give the query and documents together
  - Astute RAG: pick the best group of matching info from retrieved and internal knowledge to answer

Experiment Results

Model	Method	FaithEval	AmbigDocs	RAMDocs
Llama3.3-70B-Inst	No RAG	26.70	4.30	5.80
	Prompt-based	27.30	54.20	32.60
	Astute RAG	37.10	46.80	31.80
	MADAM-RAG	**43.10**	**58.20**	**34.40**
Qwen2.5-72B-Inst	No RAG	26.40	1.80	4.20
	Prompt-based	38.50	41.20	20.60
	Astute RAG	44.60	39.80	20.80
	MADAM-RAG	**57.70**	**52.70**	**26.40**
GPT-4o-mini	No RAG	31.00	1.00	2.50
	Prompt-based	21.00	51.50	25.00
	Astute RAG	34.00	15.00	13.00
	MADAM-RAG	**38.50**	**63.00**	**28.00**

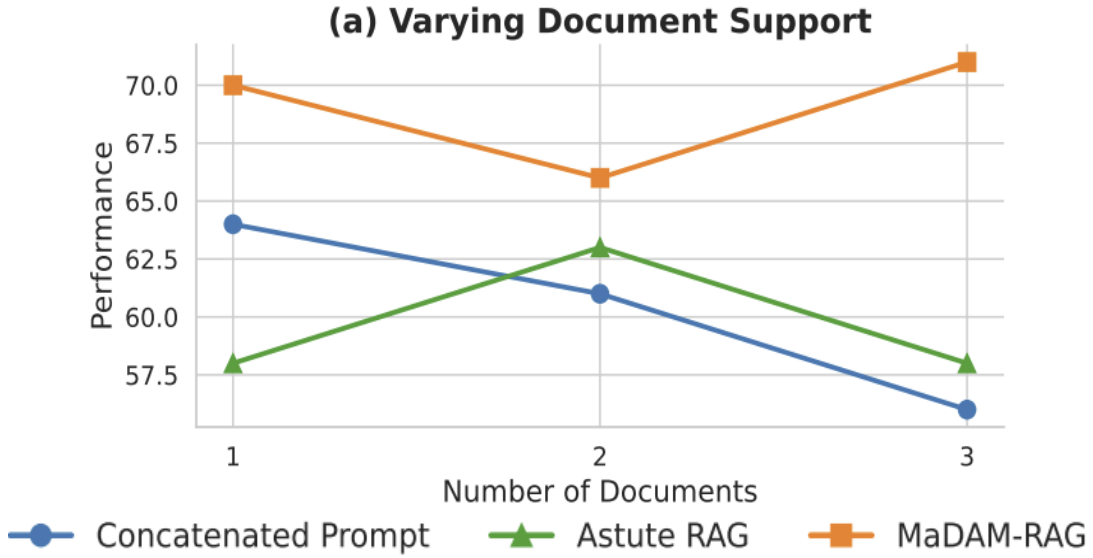
- MADAM-RAG outperforms baselines across tasks
- RAMDocs is a challenging RAG setting

Importance of Using the Aggregator and Multiple Rounds of Debate



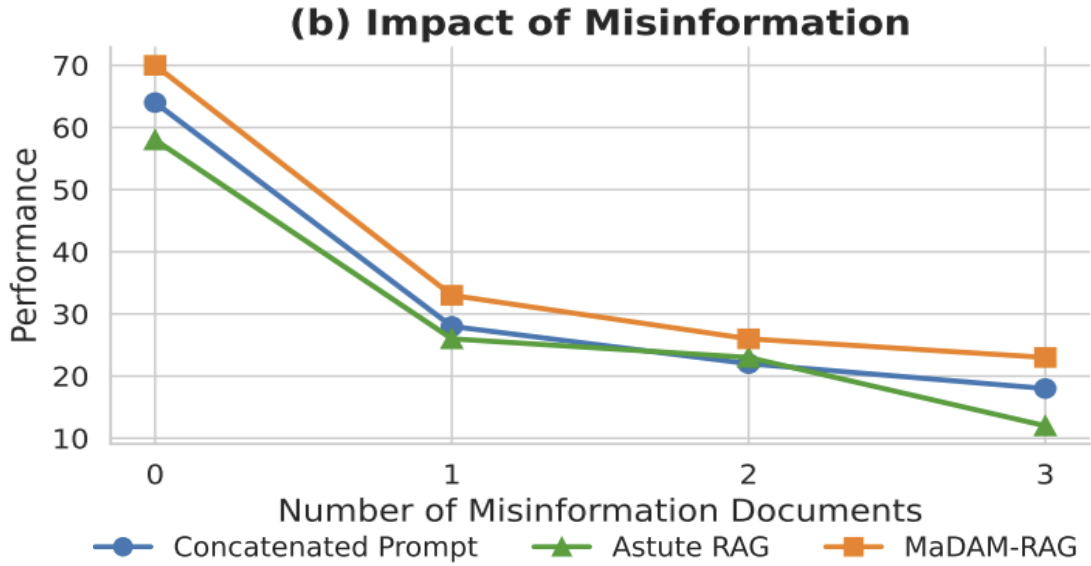
- More debate rounds and using an aggregator both help improve results
  - More rounds help agents fix mistakes and improve answers
  - The aggregator works best early by merging evidence and blocking misinformation
- In conflict settings like RAMDocs, precision matters more than recall
  - It's better to skip unsure answers than risk wrong ones

Impact of varying the number of Supporting Documents



- More supporting documents lower baseline performance
  - With imbalanced evidence, baselines prefer well-supported answers and miss others
- MADAMRAG handles imbalance well, letting one agent defend a correct answer even with less support

Impact of Increasing Misinformation



- More misinformation in the evidence leads to worse performance
- More misinformation hurts performance because it makes it harder for the LLM to trust and find the right facts
- Challenges for future work

Conclusion

- RAMDocs is a benchmark that evaluates models by considering ambiguity, conflicting answers, misinformation, and noise all together.
- Propose MADAM-RAG, where independent LLM agents debate based on individual documents and an aggregator combines their views
- MADAM-RAG improves performance on both standard and high-conflict datasets

My Review

- This paper handles all conflict types at once, which is simple but realistic and effective
- I think that independent agents keep weakly supported answers and improve them through debate is really important
- It’s promising that MADAM-RAG works well with both open and closed LLMs
- Still, it’s a bit disappointing that handling false information isn’t fully solved

Open Question

- RAMDocs is so challenging that all methods score low — what other factors should be considered beyond those proposed in the paper to improve this?