Verifiable RAG through Multi-Agent Debate
CS 510 Proposal - Research Track

Xiao Wang (xiaow4@illinois.edu)
Shen Zheng (shenz2@illinois.edu)

April 7, 2024

1 Research Question

Despite their remarkable performance in downstream tasks, Large Language Models (LLMs) exhibit limitations in addressing domain-specific or highly specialized queries. One common aspect is the generation of non-factual answers, or hallucination [HYM$^+$23]. Verifiable retrieval-augmented generation (RAG) aims to integrate LLMs' intrinsic knowledge with the external databases through retrieval with citations, which enables the users to verify the answer and makes the LLMs' output more trustworthy.

2 Approach

To address this problem, we propose verifiable retrieval-augmented generation (RAG) through multi-agent debate. Our methods include several LLM agents: Planner, Retriever, Verifier and Reflector. Their functionalities are listed as follows:

- **Planner**: The Planner is an agent that receives user questions and feedback from other agents as inputs. It formulates a plan for the Retriever to implement. This component is instrumental in decomposing complex problems into manageable subproblems for the Retriever.
- **Retriever**: The Retriever executes the plan devised by the Planner. It generates responses to user questions, including appropriate citations, thereby acting as the information retrieval component of the system.
- **Verifier**: Upon receiving the Retriever's output, the Verifier critically assesses whether the response sufficiently addresses the user's query. If not, it provides feedback to the Planner, initiating a new iteration. If the response is satisfactory, the result is passed on to the Reflector.

- **Reflector**: The Reflector critically examines each citation in the response, removing any that are irrelevant. This ensures that only pertinent citations are included in the final output.

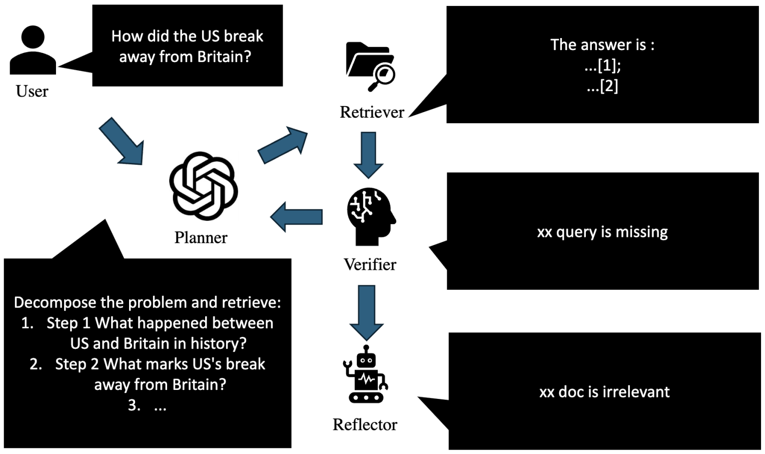
User
How did the US break away from Britain?
Retriever
The answer is :
...[1];
...[2]
Planner
Verifier
xx query is missing
Decompose the problem and retrieve:
1. Step 1 What happened between US and Britain in history?
2. Step 2 What marks US's break away from Britain?
3. ...
Reflector
xx doc is irrelevant

Figure 1: Overview

Algorithm 1 Query Resolution Process

```
Let $P$,$R$,$V$,$F$ be Planner, Retriever, Verifier, Reflector
$response \leftarrow \emptyset$
while True do
    $plan \leftarrow P(query, feedback)$
    $response \leftarrow R(plan)$
    $verificationResult \leftarrow V(response)$
    if $verificationResult$ is satisfactory then
        $filteredResponse \leftarrow F(response)$
        Output $filteredResponse$
        break
    else
        Update $feedback$ based on $verificationResult$
    end if
end while
```

3 Novelty

Current benchmarks focus on single-agent systems, with methodologies that can be categorized across three distinct dimensions:

- Problem decomposition: decompose reasoning steps obtained from the chain-of-thought (CoT) [WWS$^+$23] prompting and retrieve relevant external knowledge accordingly.
- Response refinement: update the retrieval result until it verifies that the retrieved documents are sufficient in answering the question.
- Citation refinement: identify and truncate any irrelevant citations in the final output.

Approach	Problem Decomposition	Response Refinement	Citation Refinement	Multi-agent Debate
ALCE[GYYC23]	No	No	No	No
RR[DLT$^+$23]	Yes	No	No	No
FLARE[JXG$^+$23]	No	Yes	No	No
LLatrieval[LZL$^+$24]	No	Yes	No	No
Our Method	Yes	Yes	Yes	Yes

Table 1: A summary of related work on problem decomposition, response refinement, citation refinement and multi-agent debate.

Existing work has shown that multi-agent debate can improve the LLMs' factuality and reasoning [DLT$^+$23]. Our method leverages multi-agent debate across these three dimensions, aiming to comprehensively improve the factuality and verifiability in model outputs.

4 Evaluation

We plan to evaluate on ASQA, which is a long-form factoid QA dataset where each question requires multiple short answers to cover the multiple aspects of it.

To evaluate the answer's correctness, we plan to comp[ute the recall of correct short answers by identifying whether the short answers match substrings of model output.

For the citation evaluation, we plan to uses the citation recall, citation prevision and citation F1 as metric.

5 Timeline

4/13 Dataset Collection.
4/23 Method Writing.
5/1 Experiment Implementation.
5/7 Final presentation.

6 Task Division

Xiao Wang: Responsible for report writing, experiment implementation, presentation

Shen Zheng: Responsible for report writing, experiement implementation, method design.

References

[DLT+23] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate, 2023.

[GYYC23] Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations, 2023.

[HYM+23] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, 2023.

[JXG+23] Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation, 2023.

[LZL+24] Xiaonan Li, Changtai Zhu, Linyang Li, Zhangyue Yin, Tianxiang Sun, and Xipeng Qiu. Llatrieval: Llm-verified retrieval for verifiable generation, 2024.

[WWS+23] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.