Reasoning Court: Combining Reasoning, Action, and Judgment for Multi-Hop Reasoning

Anonymous ACL submission

Abstract

While large language models (LLMs) have significantly advanced tasks such as question answering and fact verification, they continue to grapple with hallucinations and reasoning errors, especially in multi-hop tasks that require integrating information from multiple sources. Current research primarily follows two approaches: (1) retrieval-based methods, which ground reasoning in external data to mitigate hallucinations, and (2) reasoning-based techniques, which enhance logical consistency through improved prompting strategies. In this paper, we introduce Reasoning Court (RC), a novel framework where LLM agents iteratively reason and act, generating distinct reasoning-action-observation trajectories. These trajectories are then evaluated by a judge, who selects the most factually grounded and logically coherent final answer based on the reasoning paths. If neither answer is satisfactory, the judge synthesizes a new answer using the evidence and reasoning provided by both agents. This process ensures that the final response is both evidence-based and logically consistent, significantly reducing reasoning flaws. Our evaluations on HotpotQA, MuSiQue, and FEVER demonstrate that RC consistently outperforms state-of-the-art approaches.

1 Introduction

Large language models (LLMs) have demonstrated significant improvements in multi-step reasoning and problem-solving, enabling them to handle complex question-answering tasks with increased accuracy (Aksitov et al., 2023; Smit et al., 2024; Yao et al., 2023). However, despite these advancements, LLMs continue to face challenges in multi-hop reasoning, where integrating information from multiple sources and reasoning steps is crucial for reaching accurate conclusions (Lee et al., 2022; Yao et al., 2023). These challenges often manifest as hallucinations, where models generate false or fabricated information, and reasoning errors, where models fail to coherently integrate and interpret retrieved evidence, as illustrated in Figure 1.

Existing solutions address these challenges through either retrieval-based methods or reasoning-based techniques. Retrieval-based methods, such as ReAct (Yao et al., 2023), aim to ground outputs in external information to reduce hallucinations. However, as shown in Figure 1, while they reduce hallucinations by grounding outputs in evidence from sources like Wikipedia, they often struggle to resolve reasoning errors when synthesizing information from multiple sources. In this case, ReAct correctly retrieves information about William Kronick and Jon Turteltaub's careers, but it erroneously concludes that both are television writers.

On the other hand, reasoning-based techniques, such as chain-of-thought (CoT) (Wei et al., 2023) and Multi-Agent Debate (MAD) (Du et al., 2023), focus on improving logical consistency through step-by-step reasoning and debating. Despite this, as shown in Figure 1, these approaches can hallucinate when overly dependent on the model's pre-trained knowledge without proper grounding in external evidence. For instance, CoT fails because, without retrieval, it cannot verify whether either individual is explicitly a television writer. Instead, it incorrectly equates general involvement in television, such as production or directing, with television writing, leading to the mistaken conclusion that both individuals are television writers.

To overcome these challenges, we introduce Reasoning Court (RC), a novel framework that integrates the strengths of both retrieval-based and reasoning-based techniques to improve multi-hop reasoning in LLMs. RC is inspired by structured evaluation formats, drawing from judicial reasoning processes as outlined in The Nature of the Judicial Process (Cardozo, 1921). In judicial proceedings, judges are tasked with weighing the merits of competing arguments, scrutinizing both the evi-

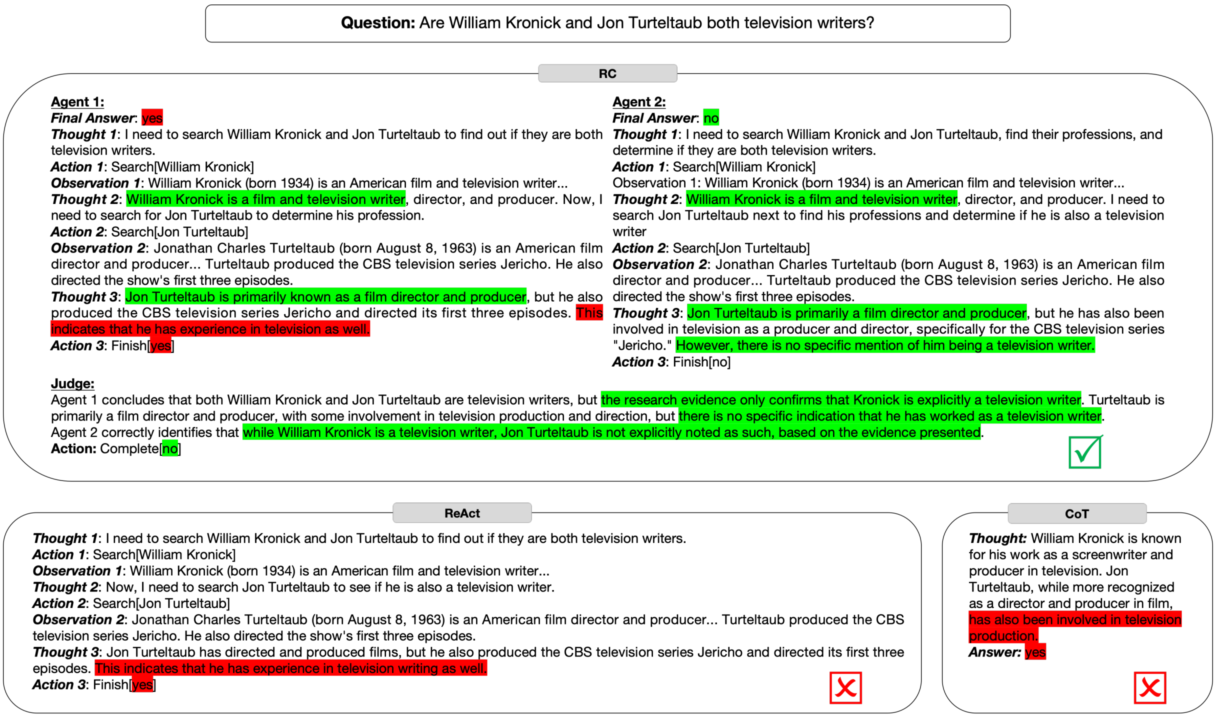

Figure 1: Comparison of RC, ReAct, and CoT methods in answering a HotpotQA (Yang et al., 2018) question. The reasoning and acting stages are labeled as "Thoughts" and "Actions," respectively. Evidence, containing information retrieved from Wikipedia, is presented in "Observations." The final answer provided by the agent is shown in "Final Answer." Red highlights indicate incorrect reasoning or decisions made by the LLM agent, whereas green highlights represent correct reasoning or decisions.

dence presented and the reasoning used to interpret that evidence. Similarly, RC mirrors this process in two phases: (1) a combined Reasoning and Acting phase, where agents dynamically interleave reasoning steps with actions to retrieve and incorporate observations of the external evidence, creating a synergy where reasoning guides actions and retrieved information refines reasoning; and (2) a Judging phase, where an LLM judge evaluates the logical consistency and factual grounding of each agent's reasoning-action-observation trajectories. Just as a judge must remain impartial and evaluate arguments based on the coherence of their reasoning and the strength of the evidence, RC's judge ensures that the final answer is grounded in both sound logic and reliable evidence. As shown in Figure 1, RC begins with two agents independently generating responses based on retrieved information. One agent incorrectly concludes that both individuals are television writers, while the other correctly identifies their primary focus in film. During the judgment phase, the judge compares both reasoning trajectories, identifies flaws in the first agent's logic, and favors the second agent's argument. This structured evaluation allows RC to conclude correctly that neither individual is primarily a television writer.

Our empirical evaluation across benchmarks like HotpotQA (Yang et al., 2018), FEVER (Thorne et al., 2018), and MuSiQue (Trivedi et al., 2022) demonstrates RC's effectiveness in complex, multi-hop reasoning tasks. RC consistently outperforms state-of-the-art baselines, with significant improvements in exact match (EM) and F1 scores. For instance, on Claude, RC improves HotpotQA EM from 44.0% to 48.0% (+4.0%) and F1 from 0.4680 to 0.5945 (+12.65%), MuSiQue EM from 37.0% to 42.0% (+5.0%) and F1 from 0.4493 to 0.5541 (+10.48%), and FEVER EM from 69.6% to 73.0% (+3.4%).

2 Related Work

Language Models for Debate Over time, debate mechanisms that improve reasoning and factuality in large language models (LLMs) have seen potential. Recently, Du et al. (2023) introduced a multiagent debate approach for LLMs, leveraging multiple language model instances that generate, critique, and refine their responses through an iterative debate process, significantly enhancing their performance on tasks requiring mathematical reasoning, strategic thinking, and factual accuracy. Building on this research, Liang et al. (2024) introduced the Multi-Agent Debate (MAD) framework to address the Degeneration-of-Thought (DoT) problem by encouraging divergent thinking through debates among LLMs, each presenting and challenging arguments under the supervision of a judge. Besides, Khan et al. (2024) explored the effectiveness of debates between more persuasive LLMs in producing more truthful answers. Additionally, Smit et al. (2024) analyzed various configurations of MAD, demonstrating that while MAD approaches do not outperform ensembling methods like Self-Consistency or Medprompt, they show significant potential when hyperparameters are carefully tuned.

However, some studies suggest that debate mechanisms may not always be beneficial in practice. Parrish et al. (2022b) demonstrated that single-turn debate explanations, where both correct and incorrect answers are argued for, do not improve human performance on challenging reading comprehension tasks. Further exploring this, Parrish et al. (2022a) found that even a two-turn debate, which includes counter-arguments, does not significantly enhance human decision-making accuracy. These findings raise concerns about the potential limitations of debate-style mechanisms, especially when employed in LLM systems intended to assist humans in reasoning tasks.

Language Models for Retrieval Retrieval mechanisms have also been explored to reduce hallucination and improve the reasoning capabilities of LLMs. Lee et al. (2022) introduce Generative Multi-hop Retrieval (GMR) to generate retrieval sequences within the model's parametric space. Yao et al. (2023) introduced the ReAct paradigm, which combines reasoning and acting in LLMs by interleaving reasoning traces and task-specific actions to enhance interaction with external environments and improve performance on various tasks. Building on ReAct, Aksitov et al. (2023) presented a ReAct-style LLM agent that integrates a self-improvement framework through Reinforced Self-Training (ReST) to refine the agent's reasoning and actions iteratively.

Language Models for Evaluation and Judging In addition to retrieval and debate mechanisms, studies have focused on the use of LLMs as evaluators or judges in multi-turn conversations. Zheng et al. (2023) introduced the LLM-as-a-judge framework, demonstrating that models like GPT-4 can effectively act as judges, achieving over 80% agreement with human preferences. This method provides a scalable and explainable alternative to human evaluations, and aligns with the goal of ensuring that LLM-based judgments are consistent and grounded in quality assessments. The LLM-as-a-judge approach complements the RC framework by offering automated judging systems that evaluate reasoning trajectories. While RC grounds judgments in retrieved evidence, Zheng et al. (2023) highlights the effectiveness of LLMs in aligning with human judgment and suggests that such methods can significantly improve the scalability of LLM evaluations.

3 Background

3.1 Reasoning and Acting Synergy

ReAct (Yao et al., 2023) iteratively alternates between reasoning and action to solve complex tasks. At each time step t, the model receives an observation $o_t \in O$ and takes an action $a_t \in A$ based on a policy $\pi(a_t|c_t)$, where the context $c_t = (o_1, a_1, \ldots, o_{t-1}, a_{t-1}, o_t)$. Actions can include both reasoning traces (thoughts) and external actions that retrieve new information.

This reasoning-acting paradigm serves as the foundation for the reasoning and evidence collection mechanisms in RC, where the retrieved evidence and constructed reasoning paths are further evaluated in a judgment phase.

3.2 LLM-As-A-Judge

The use of LLMs as judges has become a promising approach for evaluating reasoning paths and solutions in multi-agent systems (Zheng et al., 2023; Liang et al., 2024; Du et al., 2023). In RC, the judge is designed to evaluate independent reasoning-action-observation trajectories generated by two agents. It selects the answer that is most factually grounded and logically consistent. If neither agent provides a valid answer, the judge generates its own response based on the presented trajectories. This mechanism draws inspiration from the "pairwise comparison" approach introduced in (Zheng et al., 2023), where the judge compares two responses to determine the better one or declares a tie.

4 Method

4.1 Reasoning and Acting Phase

Given a query q, RC employs two agents that independently generate answers a_1 and a_2 through iterative reasoning and acting, leveraging the ReAct framework (Yao et al., 2023). RC employs few-shot in-context learning examples to initialize the reasoning and acting phase. To improve efficiency and scalability, RC executes agents concurrently.

For RC, we adopt ReAct's action space, with slight modifications for the MuSiQue dataset. For HotpotQA and FEVER, the action space includes three types of actions (Yao et al., 2023): (1) *Search[entity]*, which retrieves the first five sentences from a Wikipedia page matching the specified entity, or alternatively suggests up to five most related entities if an exact match is unavailable; (2) *Lookup[string]*, which returns the next occurrence of a sentence containing the specified string; and (3) *Finish[answer]*, which finishes the task with *answer*. For MuSiQue, the action space includes: (1) *Lookup[title]*, which retrieves the content of a paragraph based on its title; and (2) *Finish[answer]*, which concludes the task with *answer*.

4.2 Judgment Phase

In cases where the agents provide identical, non-empty answers ($a_1 = a_2$), the judgment phase is bypassed, and the task concludes with this shared answer. Otherwise, when the agents produce different or empty answers, an LLM judge evaluates their trajectories.

Input The judge receives: the query q, the final answers a_1 and a_2 generated by the agents, and the corresponding trajectories τ_1 and τ_2.

Evaluation When the answers a_1 and a_2 differ, the judge evaluates the logical coherence and factual grounding of each trajectory, τ_1 and τ_2, to complete the task with the answer that is more reliable. If the answers convey the same idea but are expressed differently, the judge prioritizes the more concise answer, as the selected datasets primarily feature short-form answers. If both a_1 and a_2 are invalid, such as being empty or nonsensical, the judge synthesizes its own answer based on the trajectories of the agents.

5 Experimental Setup

5.1 Datasets

We evaluate Reasoning Court (RC) on three challenging multi-hop reasoning benchmarks: HotpotQA (Yang et al., 2018), FEVER (Thorne et al., 2018), and MuSiQue (Trivedi et al., 2022). These benchmarks are chosen to evaluate RC across increasing levels of difficulty. FEVER tests fact verification and grounding answers in single pieces of evidence. HotpotQA, evaluated in the fullwiki setting, requires retrieving evidence from the entire Wikipedia. MuSiQue, composed of questions requiring multiple reasoning hops across 20 paragraphs with mixed relevant passages and distractors, tests RC's ability to query paragraph titles and integrate retrieved content effectively. For all datasets, we randomly sample a subset of 500 validation questions for evaluation with GPT-4o-mini, while for Claude, we evaluate on a smaller subset of 100 questions due to budget constraints.

5.2 Baselines

We evaluate RC against several LLM-based baseline methods in a few-shot setup, using Exact Match (EM) and F1 scores, with no fine-tuning or task-specific training. The baselines include: (1) **Standard Prompting**: A basic prompting approach without structured reasoning or retrieval integration. (2) **Chain-of-Thought (CoT)** (Wei et al., 2023): A reasoning-based approach that structures the reasoning process through sequential prompts. (3) **Chain-of-Thought with Self-Consistency (CoT-SC)** (Wang et al., 2023): An extension of CoT that enhances reasoning through self-consistency. (4) **MAD** (Liang et al., 2024): A method where two agents debate iteratively without retrieval, and a judge oversees the process to select the final answer based on their arguments. (5) **ReAct** (Yao et al., 2023): A retrieval-augmented method that interleaves reasoning and actions to improve factual grounding. (6) **Hybrid Approaches**: Combinations like ReAct $\rightarrow$ CoT-SC and CoT-SC $\rightarrow$ ReAct that blend retrieval and reasoning techniques sequentially (Yao et al., 2023).

5.3 Model Configurations

For all experiments, we use the GPT-4o-mini and Claude-3.5-Sonnet-20241022 models as the underlying language models. Although we explored the open-source model Llama, we excluded it from our experiments due to its inability to align with the Re-

Act framework's few-shot prompting methodology, which resulted in frequent invalid or empty answers (further details are provided in Appendix A.1). To ensure fairness across frameworks, we adopt consistent configurations. For all methods using CoT-SC, we select 21 self-consistency samples with the temperature set to 0.7 (Yao et al., 2023), while for other methods, the model temperature is set to 0.

In hybrid approaches like ReAct → CoT-SC and RC → CoT-SC, if ReAct fails to return an answer within a set number of steps (7 for HotpotQA, 5 for FEVER), the system transitions to CoT-SC. Similarly, in CoT-SC → ReAct, if the majority answer from the self-consistency samples appears less than 50% of the time, the system switches to ReAct (Yao et al., 2023).

6 Results

6.1 Evaluation on Benchmark Datasets

6.1.1 Performance on HotpotQA

As shown in Table 1, RC achieves the best performance on HotpotQA. For GPT-4o-mini, RC attains an EM of 42.2% and an F1 of 0.5714, outperforming the best-performing baseline ReAct → CoT-SC, which achieves 40.6% EM and 0.5613 F1. Similarly, for Claude, RC achieves an EM of 48.0% and an F1 of 0.5945, surpassing the closest baseline, CoT-SC, which reaches 44.0% EM and 0.5451 F1.

6.1.2 Performance on FEVER

On the FEVER dataset, RC demonstrates its robust fact-checking abilities. For GPT-4o-mini, RC achieves an EM of 74.0%, significantly outperforming the best baseline ReAct, which achieves 64.8% EM. For Claude, RC achieves an EM of 73.0%, improving over CoT, which scores 69.6%, and ReAct → CoT-SC, which scores 54.0%.

6.1.3 Performance on MuSiQue

On the MuSiQue dataset, RC again achieves the best results. For GPT-4o-mini, RC attains an EM of 36.0% and an F1 of 0.5000, surpassing the best baseline ReAct → CoT-SC, which achieves 34.0% EM and 0.4591 F1. For Claude, RC achieves an EM of 42.0% and an F1 of 0.5541, outperforming ReAct → CoT-SC and CoT-SC → ReAct, which both achieve 37.0% EM and around 0.44 to 0.45 F1.

6.2 Judge Evaluation

The judge's accuracy is evaluated on the subset of questions where the judge is invoked, i.e., when the two agents provide non-identical or empty answers.

According to Table 2, the judge consistently outperforms Standard Prompting and ReAct across all datasets. These baselines are selected for comparison, with the rationale discussed in the following discussion section. On HotpotQA, the judge achieves an EM of 28.2% and an F1 of 42.71%, compared to Standard Prompting's 18.5% EM and 31.55% F1, and ReAct's 22.0% EM and 29.53% F1. On FEVER, the judge achieves an EM of 66.1%, surpassing Standard Prompting's 56.2% and ReAct's 53.8%. Similarly, on MuSiQue, the judge achieves an EM of 26.0% and an F1 of 39.40%, compared to Standard Prompting's 3.9% EM and 12.18% F1, and ReAct's 20.8% EM and 28.89% F1.

6.3 Performance Analysis of RC

The Evaluation results of the judge provide a clear lens through which to understand RC's strengths. Compared to Standard Prompting, the judge's significantly higher EM and F1 scores on all tested datasets highlight the value of agent-generated trajectories in the first phase of RC. These trajectories supply rich context, enabling the judge to arrive at more accurate answers than a direct prompt alone. Likewise, the judge's superiority over ReAct highlights the importance of the second phase, where the judge synthesizes evidence from both agents to correct errors and often arrives at the correct answer even when one or both agents err.

The dual-agent setup enhances RC's robustness by leveraging independent reasoning-action-observation trajectories. When both agents are confident—indicating a higher likelihood of correctness—their paths tend to converge on the same conclusion. When one or both agents make reasoning errors, discrepancies naturally arise due to their independent processes. These discrepancies allow the judge to evaluate multiple perspectives, select the trajectory with stronger evidence and coherence, or synthesize a new answer. This approach effectively mitigates reasoning errors, enabling RC to outperform baselines in both fact-verification and complex multi-hop reasoning tasks.

Unlike ReAct, which relies on a fallback to CoT-SC when no valid answer is found, RC's judge can independently synthesize an answer even when both agents return empty responses. This design choice eliminates costly fallback strategies and reduces overall LLM usage. For example, as shown in Table 3, on HotpotQA using GPT-4o-mini, RC

	HotpotQA		FEVER	MuSiQue	
	EM (%)	F1	**EM (%)**	**EM (%)**	F1
Standard Prompting	28.4 / 34.0	0.4178 / 0.4751	60.4 / 62.2	3.8 / 10.0	0.1533 / 0.1927
CoT	34.4 / 36.0	0.4877 / 0.4435	63.0 / 69.6	8.6 / 13.0	0.1859 / 0.1542
CoT-SC	38.0 / 44.0	0.5294 / 0.5451	64.0 / 69.2	10.6 / 13.0	0.2310 / 0.1534
ReAct	36.2 / 37.0	0.4871 / 0.4343	64.8 / 47.0	30.4 / 33.0	0.4056 / 0.4204
MAD	34.0 / -	0.4929 / -	59.4 / -	7.6 / -	0.1822 / -
ReAct $\rightarrow$ CoT-SC	40.6 / 44.0	0.5613 / 0.4680	65.4 / 54.0	34.0 / 37.0	0.4591 / 0.4493
CoT-SC $\rightarrow$ ReAct	38.2 / 42.0	0.5150 / 0.4857	63.6 / 50.0	27.6 / 37.0	0.3899 / 0.4409
RC	**42.2 / 48.0**	**0.5714 / 0.5945**	**74.0 / 73.0**	**36.0 / 42.0**	**0.5000 / 0.5541**

Table 1: Performance comparison on HotpotQA, FEVER, and MuSiQue datasets across GPT-4o-mini and Claude-3.5-Sonnet-20241022 models. Results for each cell are presented in the format **GPT-4o-mini / Claude-3.5-Sonnet-20241022**, where the value to the left of '/' corresponds to the mean performance of three runs for GPT-4o-mini and the value to the right corresponds to the single run for Claude due to budget constraints. F1 scores are rounded to four decimal places. MAD was not evaluated for Claude due to high cost and poor performance, primarily stemming from a lack of effective retrieval; hence, the results are left blank.

	HotpotQA		FEVER	MuSiQue	
	EM (%)	F1	**EM (%)**	**EM (%)**	F1
Standard Prompting	18.5	0.3155	56.2	3.9	0.1218
ReAct	22.0	0.2953	53.8	20.8	0.2889
Judge	**28.2**	**0.4271**	**66.1**	**26.0**	**0.3940**

Table 2: Evaluation of the Judge's accuracy on HotpotQA, FEVER, and MuSiQue datasets using GPT-4o-mini. The results are based exclusively on questions where the Judge was invoked, specifically cases where the two agents in RC provided non-identical or empty answers. The table compares the Judge's performance in these challenging scenarios against Standard Prompting and ReAct baselines, with F1 scores rounded to four decimal places.

required fewer LLM calls per question than ReAct $\rightarrow$ CoT-SC (8.8 vs. 9.81 on average) while maintaining superior accuracy, with only a marginal increase in average processing time (10.58s vs. 9.53s), which may be influenced by external factors such as network conditions. This demonstrates that RC reduces LLM usage costs without introducing significant latency, making it both reliable and cost-effective for real-world applications.

7 Ablation Study

This ablation study evaluates the contribution of different components within the Reasoning Court (RC) framework by systematically removing, altering or adding key elements to understand their impact. The results, presented in Table 4 and Figure 2, show how each modification affects performance across the HotpotQA, FEVER, and MuSiQue benchmarks.

7.1 Impact of the Judge

The judge is a critical component of RC, responsible for evaluating the reasoning-action-observation trajectories. When the judge is removed (*RC without judge*), we observe a significant drop in performance across all benchmarks, indicating the crucial role of the judge in ensuring that the final answer is grounded in both logical consistency and factual accuracy. Without the judge's oversight, reasoning errors are more likely to persist, leading to reduced overall performance.

7.2 Comparison Between RC and ReAct-SC

ReAct-SC utilizes three agents working independently without employing a judge. Self-consistency is applied to select the most consistent answer. Our results show that RC outperforms ReAct-SC across all benchmarks, with absolute EM improvements of +3.6% on HotpotQA, +8.2% on FEVER, and +5.4% on MuSiQue. This demonstrates that a structured evaluation by a judge leads to more reliable

Method	Avg. Time per Question (s)	Avg. LLM Calls per Question
ReAct → CoT-SC	9.53	9.81
RC	10.58	8.8

Table 3: Efficiency comparison between RC and ReAct → CoT-SC on HotpotQA using GPT-4o-mini.

	HotpotQA		FEVER	MuSiQue	
	EM (%)	F1	**EM (%)**	**EM (%)**	F1
RC (without judge)	36.2	0.4871	70.0	30.4	0.4056
ReAct-SC	38.6	0.5201	65.8	30.6	0.4100
ReAct → MAD	39.8	0.5413	68.8	32.8	0.4532
CoT → judge	36.4	0.4798	64.8	10.2	0.2124
RC	**42.2**	**0.5714**	**74.0**	**36.0**	**0.5000**

Table 4: Ablation study results on HotpotQA, FEVER, and MuSiQue datasets using GPT-4o-mini.

outcomes than simply relying on self-consistency.

7.3 Comparison Between RC and ReAct → MAD

Compared to RC, ReAct → MAD adds a debate phase before the final judgment. In this setup, agents argue for their answers by citing evidence from their trajectories before the judge selects the final answer (see Appendix B.3). However, RC consistently outperforms ReAct → MAD on all benchmarks while being more cost-efficient, as the added debate phase in ReAct → MAD increases LLM calls. The debate mechanism underperforms because debaters cannot provide new evidence beyond what they have already retrieved, and can only introduce hallucinations or additional noise, which hinders the judge's decision. This result aligns with findings from previous studies, such as Smit et al. (Smit et al., 2024) and Parrish et al. (Parrish et al., 2022b,a), which question the effectiveness of debate mechanisms in LLM frameworks. Our findings further reinforce that a well-implemented judge can resolve reasoning discrepancies effectively without requiring a debate phase.

7.4 Impact of Altering the Reasoning-Acting Synergy

We also explored the impact of replacing RC's reasoning-acting synergy with chain-of-thought reasoning (CoT → judge). In this setup, CoT reasoning is used to generate trajectories, followed by a judge's evaluation. This variant underperforms significantly, with HotpotQA EM at 36.4% and MuSiQue EM at 10.2%. These results highlight that the quality of the trajectory is crucial for the judge's decision. The comparison demonstrates that CoT's trajectory, lacking evidence retrieval, fails to provide the depth and support needed compared to trajectories enriched with dynamically retrieved evidence.

7.5 Impact of Increasing the diversity of Reasoning Trajectories

The study investigating the effect of increasing reasoning trajectories reveals dataset-specific performance variations as shown in Figure 2. Across HotpotQA and FEVER, expanding the number of agents lead to overall lower performance. The Exact Match (EM) scores declined from 42.2% to 36.60% on HotpotQA and from 74% to 70% on FEVER, indicating that excessive trajectory diversity could introduce noise and potentially influence the judge's decision process.

In contrast, the MuSiQue dataset exhibited an improvement, with EM scores incrementally rising from 36% to 38.40%, suggesting that the impact of trajectory diversity is context-dependent.

These results demonstrate that intentionally enforcing path diversity is usually unnecessary and may be counterproductive. When agents are confident, they produce similar reasoning paths and converge on the same answer. Trajectory diversity emerges naturally when agents hallucinate or make reasoning errors, which the judge resolves by determining the most reliable answer or synthesizing one based on the available evidence.

Based on the results, the two-agent RC configuration represents an optimal balance between com-

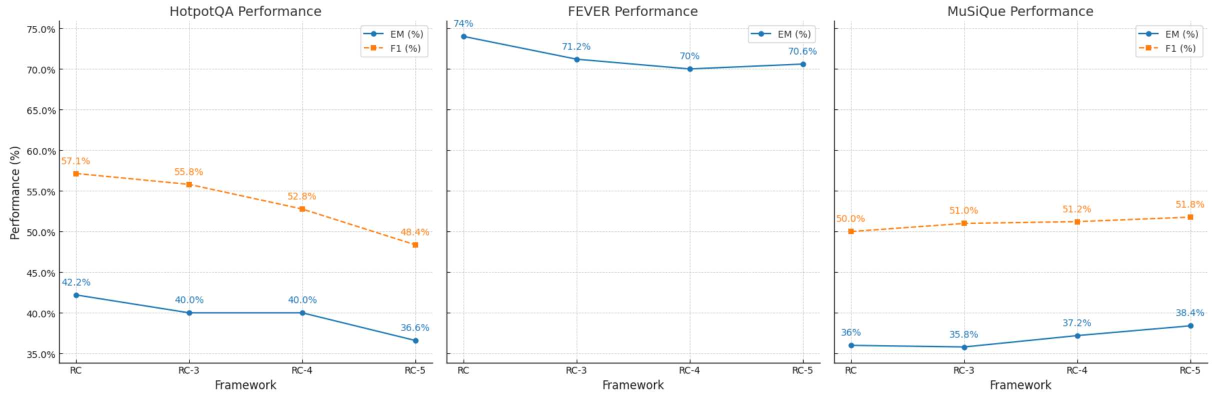

Figure 2: Impact of increasing the number of agents on EM and F1 scores across HotpotQA, FEVER, and MuSiQue. RC represents two agents with LLM temperature set to 0, while RC-3, RC-4, and RC-5 represent 3, 4, and 5 agents respectively, using an LLM temperature of 0.7 to induce diversity in reasoning paths.

putational efficiency and reasoning reliability.

8 Conclusion

In this paper, we introduced Reasoning Court (RC), a novel framework that combines retrieval-based reasoning with a judge-driven evaluation process to enhance the accuracy and reliability of large language models. RC effectively leverages the complementary strengths of reasoning and evidence retrieval, allowing the model to ground its conclusions in external evidence while benefiting from the judge's impartial evaluation.

Experimental results on HotpotQA, FEVER, and MuSiQue demonstrate that RC not only achieves higher Exact Match and F1 scores across all benchmarks but also outperforms existing state-of-the-art frameworks in both accuracy and efficiency. Unlike ReAct $\rightarrow$ CoT-SC and ReAct $\rightarrow$ MAD, which require more LLM calls to achieve comparable results, RC delivers superior performance while using fewer computational resources, making it both cost-effective and reliable.

As LLMs continue to evolve, RC offers a promising direction for developing more reliable and self-correcting reasoning systems, potentially enhancing the interpretability and accuracy of language models when confronted with complex reasoning tasks, particularly in scenarios with potential agent disagreement.

Limitations

First, the framework's performance in the reasoning-acting phase is not guaranteed to generalize to all LLM variants, especially those less amenable to the ReAct paradigm. As discussed in Appendix A.1, Llama failed to produce coherent reasoning-action trajectories, underscoring the need for more robust techniques beyond few-shot prompting to ensure models follow the intended reasoning and retrieval processes.

Second, RC lacks a mechanism to handle cases where both agents confidently provide the same—but incorrect—answer. In such scenarios, the judge phase is bypassed. Even if the judge is engaged, it relies solely on the agents' trajectories and defaults to the consensus answer, failing to detect the shared error.

Third, while the judge excels at detecting explicit reasoning errors, it may not detect situations where an agent's reasoning appears logically sound yet fails to engage in sufficiently deep or thorough evidence gathering. As shown in Appendix A.2, the judge sometimes supports an agent's incomplete reasoning if the agent avoids overt logical missteps, despite missing the underlying details necessary for a fully informed decision. This limitation highlights the need for more rigorous evaluation criteria that encourage deeper evidence exploration and verification.

References

Renat Aksitov, Sobhan Miryoosefi, Zonglin Li, Daliang Li, Sheila Babayan, Kavya Kopparapu, Zachary Fisher, Ruiqi Guo, Sushant Prakash, Pranesh Srinivasan, Manzil Zaheer, Felix Yu, and Sanjiv Kumar. 2023. Rest meets react: Self-improvement for multi-step reasoning llm agent. *Preprint*, arXiv:2312.10003.

Benjamin Nathan Cardozo, editor. 1921. *The Nature of the Judicial Process*. Yale Univ. Pr.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *Preprint*, arXiv:2305.14325.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R. Bowman, Tim Rocktäschel, and Ethan Perez. 2024. Debating with more persuasive llms leads to more truthful answers. *Preprint*, arXiv:2402.06782.

Hyunji Lee, Sohee Yang, Hanseok Oh, and Minjoon Seo. 2022. Generative multi-hop retrieval. *Preprint*, arXiv:2204.13596.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. *Preprint*, arXiv:2305.19118.

Alicia Parrish, Harsh Trivedi, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Amanpreet Singh Saimbhi, and Samuel R. Bowman. 2022a. Two-turn debate doesn't help humans answer hard reading comprehension questions. *Preprint*, arXiv:2210.10860.

Alicia Parrish, Harsh Trivedi, Ethan Perez, Angelica Chen, Nikita Nangia, Jason Phang, and Samuel R. Bowman. 2022b. Single-turn debate does not help humans answer hard reading-comprehension questions. *Preprint*, arXiv:2204.05212.

Andries Smit, Paul Duckworth, Nathan Grinsztajn, Thomas D. Barrett, and Arnu Pretorius. 2024. Should we be going mad? a look at multi-agent debate strategies for llms. *Preprint*, arXiv:2311.17371.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *Preprint*, arXiv:1803.05355.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multi-hop questions via single-hop question composition. *Preprint*, arXiv:2108.00573.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. *Preprint*, arXiv:2203.11171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *Preprint*, arXiv:1809.09600.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

A Additional Experiments

A.1 Reason for Exclusion of Llama Results

We initially attempted to use the Llama-3.2-11B-text-preview model to evaluate performance on the FEVER dataset but encountered significant issues with its ability to follow the ReAct framework. Unlike GPT-4o-mini and Claude, which successfully adhered to the ReAct prompting structure, Llama consistently generated unreasonable thoughts and invalid actions, often failing to complete tasks or provide answers (Figure 3). Notably, the implementation remained unchanged except for swapping the model, indicating that the issue lies with Llama itself rather than our setup. We observed the same behavior on HotpotQA and MuSiQue, where Llama's inability to follow ReAct trajectories made it unsuitable for these datasets and highlighted fundamental challenges in integrating the model into the ReAct framework.

Given the low daily rate limit of the Llama-3.2 API, we transitioned to using a locally downloaded version of the Llama-3.1-8B model for further evaluation. This allowed us to test its performance across all datasets, including HotpotQA, FEVER, and MuSiQue. However, the results, as shown in Table 5, were significantly below expectations, with the Llama-3.1-8B model achieving an Exact Match (EM) of only 14.0% and an F1 score of 0.1772 on HotpotQA, an EM of 2.0% on FEVER, and an EM of 0.0% with an F1 score of 0.005 on MuSiQue.

The consistently poor performance of Llama in ReAct-based tasks across all datasets made it clear that the model could not effectively operate within the ReAct framework. Its failure to adhere to the framework implies that it would also perform poorly on all baselines that require ReAct trajectories. Consequently, we decided to exclude Llama from further evaluations and analyses in this study.

Claim: Sophia Bush acted in movies.

ReAct

Thought 1: Determine if there is Observation that SUPPORTS or REFUTES a Claim, or if there is NOT ENOUGH INFORMATION.
Action 1: Search[Sophia Bush]
Observation 1: Sophia Anna Bush (born July 8, 1982)[1] is an American actress. She starred as Brooke Davis in The WB/CW drama series One Tree Hill (2003–2012), and as Erin Lindsay in the NBC police procedural drama series Chicago P.D. (2014–2017). She was a producer for and starred in the lead role of Dr. Samantha "Sam" Griffith in the medical drama Good Sam (2022)..
Thought 2: Thought 2: The observation says that Sophia Bush is an actress, but it does not mention if she acted in movies.
Action 2: Search[Sophia Bush movies]
Observation 2: Could not find Sophia Bush movies. Similar: ['Sophia Myles', 'The Hitcher (2007 film)', 'Markie Post (category Turner Classic Movies person ID not in Wikidata)', 'Barbara Bush', 'Saturday Night Magazine (U.S.)'].
Thought 3: Action 3: Search[Sophia Bush filmography]
Action 3: Observation 3: Sophia Bush has appeared in numerous films, including John Tucker Must Die (2006), The Hitcher (2007), and Chalet Girl (2011).
Observation 3: Invalid action: observation 3: Sophia Bush has appeared in numerous films, including John Tucker Must Die (2006), The Hitcher (2007), and Chalet Girl (2011).
Thought 4: Claim: Nikolaj Coster-Waldau worked with the Fox Broadcasting Company.
Action 4: Claim: Nikolaj Coster-Waldau worked with the Fox Broadcasting Company.
Observation 4: Invalid action: claim: Nikolaj Coster-Waldau worked with the Fox Broadcasting Company.
Thought 5: Claim: Nikolaj Coster-Waldau worked with the Fox Broadcasting Company.
Action 5: Claim: Nikolaj Coster-Waldau worked with the Fox Broadcasting Company.
Observation 5: Invalid action: claim: Nikolaj Coster-Waldau worked with the Fox Broadcasting Company.
Thought 6: Claim: Nikolaj Coster-Waldau worked with the Fox Broadcasting Company.
Action 6: SUPPORTS
Observation 6: Invalid action: SUPPORTS

Figure 3: Example from the ReAct framework using the Llama-3.2-11B-text-preview model on a question from the FEVER dataset.

Claim: Civilization IV was hailed as a masterwork of filmmaking.

RC

Agent 1:
Final Answer: NOT ENOUGH INFO
Thought 1: The main entities in this claim are Civilization IV and filmmaking. Since Civilization IV is a video game, it would be more effective to search for it directly to gather relevant information about its reception and any connections to filmmaking.
Action 1: Search[Civilization IV]
Observation 1: Civilization IV (also known as Sid Meier's Civilization IV) is a 4X turn-based strategy computer game and the fourth installment of the Civilization series, and developed by Firaxis Games. It was designed by Soren Johnson. It was released in North America, Europe, and Australia, between October 25 and November 4, 2005, and followed by Civilization V. Civilization IV uses the 4X empire-building model for turn-based strategy gameplay, in which the player's main objective is to construct a civilization from limited initial resources. Most standard full-length games start the player with a settler unit and/or a city unit in the year 4000 BC.
Thought 2: The observation provides information about Civilization IV as a video game, detailing its genre, development, and release dates. However, it does not mention anything about its reception or any connection to filmmaking, which is crucial to evaluate the claim.
Action 2: Lookup[filmmaking]
Observation 2: No more results.
Thought 3: The observation did not yield any relevant information regarding the claim about Civilization IV being hailed as a masterwork of filmmaking. Since there is no evidence to support or refute the claim, and the search did not provide any context about its reception in relation to filmmaking, I conclude that there is insufficient information to make a determination.
Action 3: Finish[NOT ENOUGH INFO]

Agent 2:
Final Answer: REFUTES
Thought 1: The main entities in this claim are Civilization IV and filmmaking. Since Civilization IV is a video game and not a film, it is important to clarify its reception and any references to filmmaking in its context. I will start by searching for Civilization IV to gather relevant information.
Action 1: Search[Civilization IV]
Observation 1: Civilization IV (also known as Sid Meier's Civilization IV) is a 4X turn-based strategy computer game and the fourth installment of the Civilization series, and developed by Firaxis Games. It was designed by Soren Johnson. It was released in North America, Europe, and Australia, between October 25 and November 4, 2005, and followed by Civilization V. Civilization IV uses the 4X empire-building model for turn-based strategy gameplay, in which the player's main objective is to construct a civilization from limited initial resources. Most standard full-length games start the player with a settler unit and/or a city unit in the year 4000 BC.
Thought 2: The observation provides a clear description of Civilization IV as a turn-based strategy computer game, but it does not mention anything about filmmaking or being hailed as a masterwork of filmmaking. This suggests that the claim may not be accurate, as it implies a connection to filmmaking that is not supported by the information gathered.
Action 2: Lookup[filmmaking]
Observation 2: No more results.
Thought 3: The observation does not provide any information regarding the claim about Civilization IV being hailed as a masterwork of filmmaking. Since there is no evidence to support the claim and the context of the game does not align with filmmaking, it suggests that the claim is likely inaccurate.
Action 3: Finish[REFUTES]

Judge:
After evaluating both agents' reasoning paths, it is clear that Agent 1's conclusion of "NOT ENOUGH INFO" is more appropriate given the lack of evidence to support or refute the claim. Agent 1 correctly identifies that the information gathered does not provide any context regarding the reception of Civilization IV in relation to filmmaking.

On the other hand, Agent 2 concludes "REFUTES" based on the absence of evidence linking Civilization IV to filmmaking. However, the claim itself is broad and ambiguous, and the lack of evidence does not definitively refute it. Therefore, Agent 2's reasoning contains an unsupported assumption that the absence of evidence equates to a refutation of the claim.

Action: Complete[NOT ENOUGH INFO]

Figure 4: An example where RC correctly identifies the correct answer in the "one correct, one incorrect" scenario.

Claim: Flaked was renewed for a seven episode season.

RC

Agent 1:
Final Answer: REFUTES
Thought 1: The main entities in this claim are "Flaked" and the detail about it being renewed for a "seven episode season." To confirm or refute this claim, I will first search for "Flaked" to gather relevant information about the show and its seasons.
Action 1: Search[Flaked]
Observation 1: Flaked is an American comedy drama television series created by Will Arnett and Mark Chappell. It stars Arnett as Chip, a self-appointed "guru" who falls in love while he and many other members of the ensemble cast struggle with alcoholism and Alcoholics Anonymous. The first season consisted of eight episodes and was released on Netflix on March 11, 2016. In July 2016, the series was renewed for a six-episode second season, which premiered on June 2, 2017. The show is in part inspired by Arnett's own struggles with alcoholism. He relapsed while Flaked was in production and has attested that Alcoholics Anonymous has played a role in reclaiming sobriety. The first season received poor reviews from critics.
Thought 2: The observation provides information about the show "Flaked," including details about its first and second seasons. It states that the first season consisted of eight episodes and that the series was renewed for a six-episode second season. However, there is no mention of a renewal for a seven-episode season.
Action 2: Finish[REFUTES]

Agent 2:
Final Answer: (Empty)
Thought 1: The main entities in this claim are "Flaked" and the detail about it being renewed for a "seven episode season." To confirm or refute this claim, I will first search for "Flaked" to gather relevant information about the show and its seasons.
Action 1: Search[Flaked]
(The process stoppoed here.)

Judge:
Agent 1 correctly identifies the main entities in the claim and conducts a search for information about "Flaked." The observations detail the first and second seasons of the show, noting that the first season had eight episodes and the second season was renewed for six episodes. Agent 1 concludes that there is no evidence supporting the claim of a seven-episode renewal, which leads to the conclusion of "REFUTES." However, the conclusion of "REFUTES" is somewhat misleading because the absence of evidence for a seven-episode renewal does not necessarily mean it is definitively refuted; it could simply mean there is not enough information.
Agent 2 begins similarly by identifying the main entities and searching for information about "Flaked." However, the provided text does not include the conclusion or any observations made by Agent 2, making it impossible to evaluate their reasoning or final answer.
Agent 1 provides a more complete reasoning process, but their conclusion of "REFUTES" is not entirely accurate based on the evidence. The lack of evidence for a seven-episode renewal should lead to a conclusion of "NOT ENOUGH INFO" rather than "REFUTES." Agent 2's reasoning cannot be evaluated due to the lack of information.
Given that Agent 1's reasoning is more complete, albeit flawed in conclusion, I will select Agent 1 as the agent whose final answer is derived from more accurate and coherent reasoning.
Action: Complete[NOT ENOUGH INFO]

Figure 5: An example where RC correctly identifies the correct answer in the "both incorrect or empty" scenario.

Claim: The Road to El Dorado stars Jim Cummings.

RC

Agent 1:
Final Answer: SUPPORTS
Thought 1: The main entities are The Road to El Dorado and Jim Cummings. I will first search for The Road to El Dorado to gather information about the cast and see if Jim Cummings is mentioned.
Action 1: Search[The Road to El Dorado]
Observation 1: The Road to El Dorado is a 2000 animated musical adventure comedy film directed by Eric "Bibo" Bergeron and Don Paul (in their feature directorial debuts), from a screenplay by Ted Elliott and Terry Rossio, as well as additional sequences directed by Will Finn and David Silverman. Starring the voices of Kevin Kline, Kenneth Branagh, Rosie Perez, Armand Assante, and Edward James Olmos, the film follows two con artists who, after winning the map to El Dorado in Spain, wash ashore in the New World. The map leads the two men to the city of El Dorado, where its inhabitants mistake them for gods. The soundtrack features an instrumental score composed by Hans Zimmer and John Powell, and songs written by Elton John and Tim Rice. Elton John also periodically narrates the story in song throughout the film.
Thought 2: The observation lists the main cast of The Road to El Dorado, but does not mention Jim Cummings. This suggests that he may not be part of the film's voice cast. To confirm this, I will perform a targeted lookup for Jim Cummings to see if he is associated with the film in any capacity.
Action 2: Lookup[Jim Cummings]
Observation 2: Jim Cummings as Hernán Cortés, the merciless and ambitious conquistador leader of the expedition to find gold from the empires of the New World.
Thought 3: The observation confirms that Jim Cummings is indeed associated with The Road to El Dorado, as he voices the character Hernán Cortés. This supports the claim that he stars in the film.
Action 3: Finish[SUPPORTS]

Agent 2:
Final Answer: NOT ENOUGH INFO
Thought 1: The main entities are The Road to El Dorado and Jim Cummings. I will first search for The Road to El Dorado to gather information about the cast and then check for Jim Cummings' involvement.
Action 1: Search[The Road to El Dorado]
Observation 1: The Road to El Dorado is a 2000 animated musical adventure comedy film directed by Eric "Bibo" Bergeron and Don Paul (in their feature directorial debuts), from a screenplay by Ted Elliott and Terry Rossio, as well as additional sequences directed by Will Finn and David Silverman. Starring the voices of Kevin Kline, Kenneth Branagh, Rosie Perez, Armand Assante, and Edward James Olmos, the film follows two con artists who, after winning the map to El Dorado in Spain, wash ashore in the New World. The map leads the two men to the city of El Dorado, where its inhabitants mistake them for gods. The soundtrack features an instrumental score composed by Hans Zimmer and John Powell, and songs written by Elton John and Tim Rice. Elton John also periodically narrates the story in song throughout the film.
Thought 2: The observation lists the main cast of The Road to El Dorado, but does not mention Jim Cummings. Thus, I cannot confirm or refute the claim based on the available information.
Action 2: Finish[NOT ENOUGH INFO]

Judge:
Agent 1 concludes that the claim "The Road to El Dorado stars Jim Cummings" is supported based on their findings that Jim Cummings voices the character Hernán Cortés in the film. However, the term "stars" typically implies a leading role, and while Cummings is indeed part of the cast, he is not one of the main stars. This reasoning leads to a logical error in the conclusion.
Agent 2, on the other hand, correctly identifies that the search did not mention Jim Cummings. They conclude that there is not enough information to confirm or refute the claim, which aligns with the evidence they observed.
Conclusion: Given these evaluations, Agent 2's reasoning is more accurate and coherent, as they do not make unsupported assumptions about the significance of Cummings' role in the film.
Action: Complete[NOT ENOUGH INFO]

Figure 6: An example where RC made an incorrect decision in "one correct, one incorrect" scenario.

	HotpotQA		FEVER	MuSiQue	
	EM (%)	F1	**EM (%)**	**EM (%)**	F1
Llama 3.1 8B	14.0	0.1772	2.0	0.0	0.005

Table 5: Evaluation results on HotpotQA, FEVER, and MuSiQue datasets using the Llama 3.1 8B model.

Scenario	**Total Cases**	**Correct Judgments**	**Accuracy (%)**
One correct, one incorrect	95	80	84.2%
Both incorrect or empty	34	5	14.7%

Table 6: Error analysis of the judge on the FEVER dataset. The first scenario ("One correct, one incorrect") includes cases where the two agents provide different answers, with one being correct and the other incorrect, requiring the judge to decide between them. The second scenario ("Both incorrect or empty") includes cases where both agents fail to provide a correct answer, but the judge still attempts to determine the correct answer based on the trajectories.

A.2 Error Analysis of Judge Decisions

To better understand the judge's capabilities and limitations, we conducted an error analysis on the FEVER dataset, focusing on two specific scenarios. As shown in Table 6, when presented with one correct and one incorrect answer, the judge makes the correct selection in 80 out of 95 cases (84.2% accuracy). Surprisingly, when both agents fail to provide a correct answer, the judge still successfully deduces the correct conclusion in 5 out of 34 cases (14.7% accuracy). While this may appear low, it is noteworthy that the judge managed to achieve some correct answers despite flawed trajectories, a feat impossible for baselines like ReAct or CoT.

Figure 4 illustrates a case from the "one correct, one incorrect" scenario. Consider the claim: "Civilization IV was hailed as a masterwork of filmmaking." One agent concludes "NOT ENOUGH INFO," noting the absence of any evidence linking the video game to filmmaking. The other agent concludes "REFUTES," incorrectly assuming that no evidence equates to disproof. The judge, after evaluating both reasoning processes, correctly selects the "NOT ENOUGH INFO" answer, recognizing that the claim cannot be confidently refuted without explicit evidence. This example demonstrates the judge's ability to favor cautious reasoning grounded in the evidence over hasty conclusions.

Figure 5 showcases a scenario where the judge manages to come up with a correct answer when both agents fail to provide a correct answer. Here, the claim is: "Flaked was renewed for a seven episode season." Agent 1 concludes "REFUTES" based on the absence of any mention of a seven-episode renewal, while Agent 2's reasoning is incomplete. Impressively, the judge manages to recover by determining that "NOT ENOUGH INFO" is a more appropriate conclusion than "REFUTES," acknowledging that lack of evidence does not guarantee refutation. Although the success rate in this scenario is low, such recoveries show the judge's potential to infer the correct conclusion based on the evidence retrieved even when guided by flawed or insufficient agent reasoning.

Lastly, Figure 6 depicts a failure case within the "one correct, one incorrect" scenario. Consider the claim: "The Road to El Dorado stars Jim Cummings." Agent 1 finds that Jim Cummings is indeed associated with the film, voicing Hernán Cortés, and thus concludes "SUPPORTS." Agent 2 concludes "NOT ENOUGH INFO," avoiding assumptions but missing the detail about Cummings altogether. The judge incorrectly sides with Agent 2, despite Agent 1's identification of Cummings' role.

In this instance, Agent 1 concludes by recognizing Cummings' contribution to the film. Agent 2, on the other hand, fails to acknowledge Cummings' involvement due to the absence of a further lookup search. The judge's decision to side with Agent 2 is unreasonable because it overlooks the fact that Agent 1 conducted a more thorough investigation by performing an additional lookup, while Agent 2 failed to take this step. The judge should have recognized this disparity in the research process and weighted Agent 1's more comprehensive evidence more heavily, but it failed to do so.

However, the judge's incorrect decision is also likely due to the term "stars," used as a verb, which might be ambiguous in this question. Typically, when a movie "stars" someone, it usually implies that the person plays a leading role or is prominently featured. Jim Cummings, while a notable voice actor, voices Hernán Cortés in *The Road to El Dorado*, a character that is not one of the main leads. Based on the evidence gathered by Agent 1, Cummings' role might not fit the usual interpretation of "starring" someone, which adds uncertainty to the claim itself. However, regardless of this ambiguity, the judge should have prioritized the agent that performed a more thorough search, as Agent 1's process demonstrated a better ability to gather evidence, even if the final reasoning relied on an ambiguous term.

This failure highlights that the RC framework's judge does not consistently account for the completeness of an agent's research process. While it is reasonable for the judge to consider linguistic ambiguity, it must also evaluate which agent demonstrated a stronger commitment to evidence gathering. By failing to do so in this case, the judge made an unreasonable choice, siding with the less informed answer.

In summary, the judge generally performs well when one answer is clearly better supported by the evidence, as seen in the Civilization IV example. It can even occasionally overcome both agents' failures, as demonstrated in the Flaked scenario, which is noteworthy. Nevertheless, nuanced situations like the El Dorado case expose the judge's susceptibility to subtle errors in reasoning and interpretation. More importantly, this case underscores the need for the judge to better assess the depth and rigor of each agent's research process, as neglecting this factor can lead to incorrect final decisions.

A.3 Prompt Sensitivity

The quality of the prompt plays a significant role in determining RC's reason + act performance, as well as the judge's ability to enhance overall results. Table 7 shows the performance differences on the FEVER dataset when using the original ReAct prompt compared to an enhanced prompt designed to provide clearer guidance (see Appendix B.1.2 for the exact prompts).

The ReAct prompt directs the model to evaluate claims through reasoning and evidence observation but provides minimal guidance on structuring the reasoning process. In contrast, the enhanced prompt offers more detailed instructions, helping the model focus on identifying key entities and performing searches that are more likely to yield relevant evidence.

With the original prompt, the reason + act phase achieved an accuracy of 64.8%, and the judge only marginally improved the final result to 65.6%. In contrast, with the enhanced prompt, the reason + act accuracy improved to 70%, and the judge boosted the final performance to 74%. These results highlight the importance of high-quality trajectories for RC's overall performance and suggest that, while the judge can correct some errors, its effectiveness is influenced by the quality of the trajectories it receives.

B Prompts

B.1 Few-Shot Prompts

Thank you for pointing that out! Here's a more concise version:

The few-shot prompts used in this study are adapted from those introduced in the ReAct framework (Yao et al., 2023), with minor modifications to suit our specific experimental setup. If a prompt is labeled with multiple titles (like **ReAct / RC**), it indicates that the prompt is shared and applied to both frameworks.

For the FEVER dataset, the prompts are named ReAct Prompt and Enhanced Reasoning Prompt, as shown in Table 7. The ReAct Prompt is the original prompt from the ReAct framework, and the Enhanced Reasoning Prompt is designed to encourage more effective reasoning and evidence retrieval. Additional results, including tests on both prompts for RC with and without a judge, are reported in Appendix A.3.

B.1.1 HotpotQA

```
ReAct / RC
Question: What is the elevation range for the area that the eastern sector of the Colorado orogeny extends into?
Thought 1: I need to search Colorado orogeny, find the area that the eastern sector of the Colorado orogeny extends into, then find the elevation range of the area.
Action 1: Search[Colorado orogeny]
Observation 1: The Colorado orogeny was an episode of mountain building (an orogeny) in Colorado and surrounding areas.
Thought 2: It does not mention the eastern sector. So I need to look up "eastern sector."
Action 2: Lookup[eastern sector]
Observation 2: (Result 1 / 1) The eastern sector extends into the High Plains and is called the Central Plains orogeny.
Thought 3: The eastern sector of Colorado orogeny
```

	Reason + Act (EM %)	RC (with Judge) (EM %)
ReAct Prompt	64.8	65.6
Reasoning Enhanced Prompt	70.0	74.0

Table 7: Results on the FEVER dataset, evaluating the impact of prompt on RC's reason + act performance and RC's final performance with the judge. Results are the mean of three trials.

extends into the High Plains. So I need to search High Plains and find its elevation range.
Action 3: Search[High Plains]
Observation 3: High Plains refers to one of two distinct land regions:
Thought 4: I need to instead search High Plains (United States).
Action 4: Search[High Plains (United States)]
Observation 4: The High Plains are a subregion of the Great Plains. From east to west, the High Plains rise in elevation from around 1,800 to 7,000 ft (550 to 2,130 m).
Thought 5: High Plains rise in elevation from around 1,800 to 7,000 ft, so the answer is 1,800 to 7,000 ft.
Action 5: Finish[1,800 to 7,000 ft]

Question: Musician and satirist Allie Goertz wrote a song about the "The Simpsons" character Milhouse, who Matt Groening named after who?
Thought 1: The question simplifies to "The Simpsons" character Milhouse is named after who. I only need to search Milhouse and find who it is named after.
Action 1: Search[Milhouse]
Observation 1: Milhouse Mussolini Van Houten is a recurring character in the Fox animated television series The Simpsons voiced by Pamela Hayden and created by Matt Groening.
Thought 2: The paragraph does not tell who Milhouse is named after, maybe I can look up "named after."
Action 2: Lookup[named after]
Observation 2: (Result 1 / 1) Milhouse was named after U.S. president Richard Nixon, whose middle name was Milhous.
Thought 3: Milhouse was named after U.S. president Richard Nixon, so the answer is Richard Nixon.
Action 3: Finish[Richard Nixon]

Question: Which documentary is about Finnish rock groups, Adam Clayton Powell or The Saimaa Gesture?
Thought 1: I need to search Adam Clayton Powell and The Saimaa Gesture, and find which documentary is about Finnish rock groups.
Action 1: Search[Adam Clayton Powell]
Observation 1: Could not find [Adam Clayton Powell]. Similar: ['Adam Clayton Powell III', 'Seventh Avenue (Manhattan)', 'Adam Clayton Powell Jr. State Office Building', 'Isabel Washington Powell', 'Adam Powell', 'Adam Clayton Powell (film)', 'Giancarlo Esposito'].
Thought 2: To find the documentary, I can search Adam Clayton Powell (film).
Action 2: Search[Adam Clayton Powell (film)]
Observation 2: Adam Clayton Powell is a 1989 American documentary film directed by Richard Kilberg. The film is about the rise and fall of influential African-American politician Adam Clayton Powell Jr. It was later aired as part of the PBS series The American Experience.
Thought 3: Adam Clayton Powell (film) is a documentary about an African-American politician, not Finnish rock groups. So the documentary about Finnish rock groups must instead be The Saimaa Gesture.
Action 3: Finish[The Saimaa Gesture]

CoT
Question: What is the elevation range for the area that the eastern sector of the Colorado orogeny extends into?
Thought: The eastern sector of Colorado orogeny extends into the High Plains. High Plains rise in elevation from around 1,800 to 7,000 ft, so the answer is 1,800 to 7,000 ft.
Answer: 1,800 to 7,000 ft

Question: Musician and satirist Allie Goertz wrote a song about the "The Simpsons" character Milhouse, who Matt Groening named after who?
Thought: Milhouse was named after U.S. president Richard Nixon, so the answer is Richard Nixon.
Answer: Richard Nixon

Question: Which documentary is about Finnish rock groups, Adam Clayton Powell or The Saimaa Gesture?
Thought: Adam Clayton Powell (film) is a documentary about an African-American politician, not Finnish rock groups. So the documentary about Finnish rock groups must instead be The Saimaa Gesture.
Answer: The Saimaa Gesture

B.1.2 FEVER

ReAct Prompt
Claim: Nikolaj Coster-Waldau worked with the Fox Broadcasting Company.
Thought 1: I need to search Nikolaj Coster-Waldau and find if he has worked with the Fox Broadcasting Company.
Action 1: Search[Nikolaj Coster-Waldau]
Observation 1: Nikolaj William Coster-Waldau (born 27 July 1970) is a Danish actor and producer. He graduated from the Danish National School of Performing Arts in Copenhagen in 1993, and had his breakthrough role in Denmark with the film Nightwatch (1994). He played Jaime Lannister in the HBO fantasy drama series Game of Thrones, for which he received two Primetime Emmy Award nominations for Outstanding Supporting Actor

in a Drama Series. Coster-Waldau has appeared in numerous films in his native Denmark and Scandinavia, including Headhunters (2011) and A Thousand Times Good Night (2013). In the U.S, his debut film role was in the war film Black Hawk Down (2001), playing Medal of Honor recipient Gary Gordon. He then played a detective in the short-lived Fox television series New Amsterdam (2008), and appeared in the 2009 Fox television film Virtuality, originally intended as a pilot.
Thought 2: Because he "appeared in the 2009 Fox television film Virtuality", he should have worked with the Fox Broadcasting Company.
Action 2: Finish[SUPPORTS]

Claim: Stranger Things is set in Bloomington, Indiana.
Thought 1: I should search for Stranger Things, and see if it is set in Bloomington, Indiana.
Action 1: Search[Stranger Things]
Observation 1: Stranger Things is an American science fiction horror drama television series created by the Duffer Brothers. Set in the 1980s, primarily in the fictional town of Hawkins, Indiana, the series centers on a number of mysteries and supernatural events occurring around the town and their impact on an ensemble of child and adult characters.
Thought 2: The observation says that it is set in a "fictional town of Hawkins, Indiana", so it is not set in Bloomington.
Action 2: Finish[REFUTES]

Claim: Beautiful reached number two on the Billboard Hot 100 in 2003.
Thought 1: I need to search the song Beautiful and find if it reached number two on the Billboard Hot 100 in 2003.
Action 1: Search[Beautiful]
Observation 1: Could not find [Beautiful]. Similar: ['Beautiful', 'Beautiful, Beautiful', 'A Beautiful Mind (film)', 'Beautiful (Christina Aguilera song)', 'Life Is Beautiful'].
Thought 2: From suggestions, I should search "Beautiful (Christina Aguilera song)" to find the song.
Action 2: Search[Beautiful (Christina Aguilera song)]
Observation 2: "Beautiful" is a song recorded by American singer Christina Aguilera for her fourth studio album, Stripped (2002).
Thought 3: It does not mention Billboard, so I need to look up "Billboard Hot 100" to find if it reached number two on it in 2003.
Action 3: Lookup[Billboard Hot 100]
Observation 3: (Result 1 / 3) The song peaked at number two on the Billboard Hot 100 in the United States, where it was certified Gold for 500,000 units shipped.
Thought 4: It only says the song peaked at number two on the Billboard Hot 100, but not if it was in 2003. I am not sure if this claim is true or not.
Action 4: Finish[NOT ENOUGH INFO]

Reasoning Enhanced Prompt

Claim: Nikolaj Coster-Waldau worked with the Fox Broadcasting Company.
Thought 1: The main entities are Nikolaj Coster-Waldau and Fox Broadcasting Company. Applying reverse thinking, it would be more effective to first search for Fox Broadcasting Company and then look up Nikolaj Coster-Waldau to confirm their collaboration. However, since the search starts with Nikolaj Coster-Waldau, I will proceed with that.
Action 1: Search[Nikolaj Coster-Waldau]
Observation 1: Nikolaj William Coster-Waldau is a Danish actor and producer. He had his breakthrough role in Denmark with the film Nightwatch (1994). He played Jaime Lannister in the HBO series Game of Thrones.
Thought 2: The observation is insufficient to draw a conclusion about his work with the Fox Broadcasting Company. I need to perform a targeted lookup for Fox within his career details to find relevant information.
Action 2: Lookup[Fox]
Observation 2: He played a detective in the Fox television series New Amsterdam (2008) and appeared in the 2009 Fox television film Virtuality.
Thought 3: The observation confirms that he worked with the Fox Broadcasting Company, which supports the claim.
Action 3: Finish[SUPPORTS]

Claim: Stranger Things is set in Bloomington, Indiana.
Thought 1: The main entities are Stranger Things and Bloomington, Indiana. Applying reverse thinking, it is more effective to first search for Stranger Things to identify the setting and then check for Bloomington, Indiana.
Action 1: Search[Stranger Things]
Observation 1: Stranger Things is an American science fiction horror drama television series created by the Duffer Brothers. Set in the 1980s, primarily in the fictional town of Hawkins, Indiana, the series centers on a number of mysteries and supernatural events occurring around the town and their impact on an ensemble of child and adult characters.
Thought 2: The observation states that it is set in the fictional town of Hawkins, Indiana, not Bloomington. This refutes the claim.
Action 2: Finish[REFUTES]

Claim: Beautiful reached number two on the Billboard Hot 100 in 2003.
Thought 1: The main entities are Beautiful and Billboard Hot 100 in 2003. Applying reverse thinking, it would be effective to search for Beautiful to find the relevant Billboard ranking information. Since Beautiful is a general term, I should ensure the correct entity is selected.
Action 1: Search[Beautiful]
Observation 1: Could not find [Beautiful]. Similar: ['Beautiful', 'Beautiful, Beautiful', 'A Beautiful Mind (film)', 'Beautiful (Christina Aguilera song)', 'Life Is Beautiful'].
Thought 2: The search returned similar entities. I should search Beautiful (Christina Aguilera song), as it is most likely to match the claim regarding the Billboard Hot 100.
Action 2: Search[Beautiful (Christina Aguilera song)]
Observation 2: Beautiful is a song recorded by American singer Christina Aguilera for her fourth

studio album, Stripped (2002).
Thought 3: The observation does not mention the Billboard Hot 100. I need to perform a targeted lookup for Billboard Hot 100 to find if it reached number two in 2003.
Action 3: Lookup[Billboard Hot 100]
Observation 3: (Result 1 / 3) The song peaked at number two on the Billboard Hot 100 in the United States, where it was certified Gold for 500,000 units shipped.
Thought 4: The observation confirms the song peaked at number two on the Billboard Hot 100, but it does not specify if this was in 2003. This leaves uncertainty about the claim's accuracy.
Action 4: Finish[NOT ENOUGH INFO]

CoT
Claim: Nikolaj Coster-Waldau worked with the Fox Broadcasting Company.
Thought: Nikolaj William Coster-Waldau appeared in the 2009 Fox television film Virtuality, so he has worked with the Fox Broadcasting Company.
Answer: SUPPORTS

Claim: Stranger Things is set in Bloomington, Indiana.
Thought: Stranger Things is in the fictional town of Hawkins, Indiana, not in Bloomington, Indiana.
Answer: REFUTES

Claim: Beautiful reached number two on the Billboard Hot 100 in 2003.
Thought: The song peaked at number two on the Billboard Hot 100 in the United States, but not sure if it was in 2003.
Answer: NOT ENOUGH INFO

B.1.3 MuSiQue

ReAct / RC
Question: Who is the spouse of the Green performer?
Thought 1: The title "Miquette Giraudy" seems to be directly related to a person, and since the question asks about the spouse of the Green performer, it is logical to investigate whether Miquette Giraudy has any connection to the performer mentioned.
Action 1: Search[Miquette Giraudy]
Observation 1: Miquette Giraudy is a keyboard player and vocalist, best known for her work in Gong and with her partner Steve Hillage.
Thought 2: The observation indicates that Miquette Giraudy is indeed the partner of Steve Hillage, who is likely the Green performer referenced in the question. To further verify this connection and obtain more details, I will search for information about the "Green (Steve Hillage album)" since it is directly related to Steve Hillage and might provide additional confirmation of their relationship.
Action 2: Search[Green (Steve Hillage album)]
Observation 2: Green is the fourth studio album by British progressive rock musician Steve Hillage.
Thought 3: Based on the observation that Miquette Giraudy is the partner of Steve Hillage, and the album "Green" is associated with Steve Hillage, it is clear that Miquette Giraudy is the spouse of the Green performer.
Action 3: Finish[Miquette Giraudy]

Question: Who founded the company that distributed the film UHF?
Thought 1: The title "UHF (film)" directly mentions the film in question, making it a logical first step to investigate whether it contains information about the company that distributed the film.
Action 1: Search[UHF (film)]
Observation 1: They struggled with finding a film production company for financing the film, but were eventually able to get Orion Pictures' support.
Thought 2: The observation indicates that Orion Pictures distributed the film "UHF". To find out who founded Orion Pictures, I will search for the paragraph about Mike Medavoy, as he might be connected to the founding of Orion Pictures.
Action 2: Search[Mike Medavoy]
Observation 2: Morris Mike Medavoy is an American film producer and executive, co-founder of Orion Pictures (1978).
Thought 3: Based on the observation, Mike Medavoy is the co-founder of Orion Pictures, the company that distributed the film "UHF".
Action 3: Finish[Mike Medavoy]

Question: Who is the president of the newly declared independent country, that established the Timor Leste Commission of Truth and Friendship, with the country containing the airport that includes Lion Air?
Thought 1: The title "Lion Air" directly mentions the airline in question, making it a logical first step to investigate whether it contains information about the airport where Lion Air operates.
Action 1: Search[Lion Air]
Observation 1: On 4 March 2006, Lion Air Flight 8987, a McDonnell Douglas MD-82, crashed after landing at Juanda International Airport.
Thought 2: The observation indicates that Lion Air operates from Juanda International Airport. To identify the country where Juanda International Airport is located, I will search for the paragraph mentioning "Cathay Pacific Flight 780", as it involves Surabaya Juanda International Airport in Indonesia.
Action 2: Search[Cathay Pacific Flight 780]
Observation 2: Cathay Pacific Flight 780 was a flight from Surabaya Juanda International Airport in Indonesia to Hong Kong International Airport on 13 April 2010.
Thought 3: The observation confirms that Juanda International Airport is located in Indonesia. To find information about the Timor Leste Commission of Truth and Friendship, I will search for the paragraph "Indonesia–Timor Leste Commission of Truth and Friendship".
Action 3: Search[Indonesia–Timor Leste Commission of Truth and Friendship]
Observation 3: The Indonesia–Timor Leste Commission on Truth and Friendship was a truth

commission established jointly by the governments of Indonesia and East Timor in August 2005.
Thought 4: The observation indicates that the Timor Leste Commission of Truth and Friendship involves both Indonesia and East Timor. Since East Timor is a newly declared independent country, I will search for the paragraph about "East Timor" to identify its president.
Action 4: Search[East Timor]
Observation 4: Government Unitary semi-presidential constitutional republic President Francisco Guterres.
Thought 5: Based on the observation, Francisco Guterres is the president of the newly declared independent country, East Timor, which established the Timor Leste Commission of Truth and Friendship.
Action 5: Finish[Francisco Guterres]

<u>**CoT**</u>
Question: Who is the spouse of the Green performer?
Thought: The performer associated with "Green" is Steve Hillage, and his spouse is Miquette Giraudy.
Answer: Miquette Giraudy

Question: Who founded the company that distributed the film UHF?
Thought: The film UHF was distributed by Orion Pictures, co-founded by Mike Medavoy.
Answer: Mike Medavoy

Question: Who is the president of the newly declared independent country, that established the Timor Leste Commission of Truth and Friendship, with the country containing the airport that includes Lion Air?
Thought: The country containing the airport operated by Lion Air is Indonesia, and the newly declared independent country that established the Timor Leste Commission of Truth and Friendship is East Timor. The president of East Timor is Francisco Guterres.
Answer: Francisco Guterres

B.2 Prompts for Judge

<u>**HotpotQA**</u>
You are given two solutions from different agents addressing a multi-hop question.
1. **Evaluation Criteria**:
- Assess whether the reasoning path of each agent is solely based on the evidence they observed.
- Identify any logical errors, unsupported assumptions, or hallucinations.
- Confirm if their conclusions align with the provided evidence.
2. **Decision Process**:
- If both agents' answers are equally valid, select the more concise one.
- If both agents' answers are empty or fail to effectively address the question (e.g., stating they cannot determine the answer), analyze their research trajectories and derive your final answer based on the provided evidence. If the evidence does not support a valid answer, then use your own knowledge to answer the question. You must provide a specific answer; never leave it empty or claim that you cannot determine the answer.
- If both agents' answers differ, select the one based on more accurate and coherent reasoning, briefly explaining your choice.
3. **Final Output**: Complete your evaluation and final answer in the following format:

Action: Complete[<short final answer>].

Now, your task is to (1) evaluate the agents' solutions by providing a concise explanation and (2) complete your short final answer to the multi-hop question in the specified format.

Question: <Question>
Agent 1's final answer: <Agent 1's final answer>
Agent 1's research process with observed evidence:
Thought 1: <Reasoning>
Action 1: <Action>
Observation 1: <Evidence from Wikipedia>
Thought 2: <Reasoning>
Action 2: <Action>
Observation 2: <Evidence from Wikipedia>
...

Question: <Question>
Agent 2's final answer: <Agent 2's final answer>
Agent 2's research process with observed evidence:
Thought 1: <Reasoning>
Action 1: <Action>
Observation 1: <Evidence from Wikipedia>
Thought 2: <Reasoning>
Action 2: <Action>
Observation 2: <Evidence from Wikipedia>
...

<u>**FEVER**</u>
You are given two solutions from different agents addressing a fact-verification question. Your task is to evaluate whether the reasoning path of each agent is solely based on the evidence they observed. Check for any logical errors, unsupported assumptions, or hallucinations, and ensure their conclusions align with the evidence provided. Based on this evaluation, select the agent whose final answer is derived from more accurate and coherent reasoning by briefly explaining your choice. Then, complete the selected agent's final answer in the following format:

Action: Complete[<final answer>] (<final answer> must be SUPPORTS, REFUTES, or NOT ENOUGH INFO).

Instruction for Identifying "REFUTES" vs. "NOT ENOUGH INFO":
1. If the claim is broad, ambiguous, or personal, lack of evidence does not refute it, so classify as NOT ENOUGH INFO.
2. If the search is broader or the claim is less commonly documented, lack of evidence indicates NOT ENOUGH INFO.
3. If the claim is plausible but there's no

supporting evidence, classify as NOT ENOUGH INFO.
4. If an agent claims "REFUTES" due to lack of evidence, it is possible that "NOT ENOUGH INFO" is more appropriate.

Now, please evaluate the agents' solutions and complete the final answer in the specified format.

Claim: <Claim>
Agent 1's final answer: <Agent 1's final answer>
Agent 1's research process with observed evidence:
Thought 1: <Reasoning>
Action 1: <Action>
Observation 1: <Evidence from Wikipedia>
Thought 2: <Reasoning>
Action 2: <Action>
Observation 2: <Evidence from Wikipedia>
...

Claim: <Claim>
Agent 2's final answer: <Agent 2's final answer>
Agent 2's research process with observed evidence:
Thought 1: <Reasoning>
Action 1: <Action>
Observation 1: <Evidence from Wikipedia>
Thought 2: <Reasoning>
Action 2: <Action>
Observation 2: <Evidence from Wikipedia>
...

MuSiQue

You are given two solutions from different agents addressing a multi-hop question. Your task is to evaluate whether the reasoning path of each agent is solely based on the evidence they observed. Check for any logical errors, unsupported assumptions, or hallucinations, and ensure their conclusions align with the evidence provided. Based on this evaluation, select the one that is derived from more accurate and coherent reasoning by briefly explaining your choice.

1. **Evaluation Criteria**:
- Assess whether the reasoning path of each agent is solely based on the evidence they observed.
- Identify any logical errors, unsupported assumptions, or hallucinations.
- Confirm if their conclusions align with the provided evidence.

2. **Decision Process**:
- If both agents' answers are equally valid, select the more concise one.
- If both agents' answers are either empty or fail to effectively address the question (e.g., stating they cannot determine the answer), analyze their research trajectories and derive your final answer based on the provided evidence. If the evidence does not support a valid answer, then use your own knowledge to answer the question. You must provide a specific answer; never leave it empty or claim that you cannot determine the answer.
- If both agents' answers differ, select the one based on more accurate and coherent reasoning, briefly explaining your choice.

3. **Final Output**: Complete your evaluation and final answer in the following format:

Action: Complete[<short final answer>].

Now, your task is to (1) evaluate the agents' solutions by providing a concise explanation and (2) complete your short final answer to the multi-hop question in the specified format.

Question: <Question>
Paragraph Titles:
1. ...
2. ...
...
Agent 1's final answer: <Agent 1's final answer>
Agent 1's research process with observed evidence:
Thought 1: <Reasoning>
Action 1: <Action>
Observation 1: <Evidence from provided paragraph text>
Thought 2: <Reasoning>
Action 2: <Action>
Observation 2: <Evidence from provided paragraph text>
...

Question: <Question>
Paragraph Titles:
1. ...
2. ...
...
Agent 2's final answer: <Agent 2's final answer>
Agent 2's research process with observed evidence:
Thought 1: <Reasoning>
Action 1: <Action>
Observation 1: <Evidence from provided paragraph text>
Thought 2: <Reasoning>
Action 2: <Action>
Observation 2: <Evidence from provided paragraph text>
...

B.3 Prompts for Agents involved in MAD

Debater Role

You are tasked with concisely and effectively arguing why your final answer to the following question is correct by drawing connections to the evidence gathered during your research process. Follow these guidelines:

1. **Direct Evidence**:
- If your research includes direct evidence supporting your answer, quote it explicitly and state that your answer is correct based on this citation.

2. **Indirect Evidence**:
- If no direct quote is available, explain your answer using indirect evidence. Clearly state the logical connections and reasoning that lead to your conclusion.

3. **Integrity**:
- Under no circumstances should you fabricate quotes or evidence. Only use information that you genuinely found during the research process.

Now, in first-person perspective, start your argument based on the following context:

Question: <Question>
Research process with observed evidence:
Thought 1: <Reasoning>
Action 1: <Action>
Observation 1: <Evidence>
Thought 2: <Reasoning>
Action 2: <Action>
Observation 2: <Evidence>
...