Tool-MAD: Dynamic Multi-Agent Debate Framework with Adaptive Retrieval for Accurate and Hallucination-Resistant Fact Verification

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have achieved impressive results across a wide range of language tasks, but they still struggle with hallucinations and factual inaccuracies, particularly in complex reasoning and fact verification tasks. To address these limitations, we introduce Tool-MAD, a novel multi-agent debate (MAD) framework designed to enhance factual verification by equipping agents with external tools, including search APIs and Retrieval-Augmented Generation (RAG) modules. Tool-MAD incorporates two core innovations: (1) an adaptive query formulation mechanism that enables agents to iteratively refine evidence retrieval based on evolving debate contexts and prior arguments, and (2) a novel consistency score, which quantitatively assesses the semantic similarity between agents' responses and retrieved evidence, allowing the Judge agent to reliably detect hallucinations and improve factual alignment. Experimental results on four benchmark datasets for fact verification demonstrate that Tool-MAD consistently outperforms other multi-agent debate frameworks. Furthermore, in the medical question answering domain, Tool-MAD demonstrates strong robustness and flexibility across alternative tools and domain settings.

1 Introduction

Large Language Models (LLMs) have demonstrated strong performance across a wide range of natural language processing tasks in recent years (Brown et al., 2020; Thoppilan et al., 2022; Driess et al., 2023). In particular, their ability to generate fluent and coherent text has made them widely applicable to tasks such as dialogue generation, content summarization, and knowledge extraction. Despite these capabilities, LLMs frequently suffer from hallucination, which refers to generating confident yet factually incorrect information (Li et al., 2024).

To address these limitations, recent studies have proposed single-agent prompt-based methods, such as Chain-of-Thought (CoT) (Wei et al., 2022; Kojima et al., 2022), self-reflection (Shinn et al., 2023a). These approaches guide a single LLM agent through intermediate reasoning steps or enable it to leverage external resources, such as the Wikipedia API, for tool-augmented reasoning. A key advantage of these single-agent methods is their simplicity, as they require minimal changes to existing model architectures or training routines.

Nevertheless, single-agent approaches often keep repeating the same wrong answers without correcting themselves, which is known as *Degeneration of Thought* (Liang et al., 2024). To address this problem, the Multi-Agent Debate (MAD) framework has recently been proposed. By enabling multiple LLM agents to engage in dialogue with one another, MAD promotes diverse reasoning paths and fosters cross-agent verification, thereby improving answer accuracy(Du et al., 2023; Liang et al., 2024). However, existing MAD frameworks remain limited in two key aspects. First, they typically rely on the model's internal knowledge or static documents, without the capacity to actively retrieve or interact with external sources of information. Second, when a dedicated Judge agent is employed to make the final decision, its judgment is still based solely on LLM-generated content—such as the debate history—making the outcome vulnerable to the same hallucination risks present in single-agent settings.

To address the limitations of traditional MAD approaches, which solely rely on internal model knowledge, Wang et al. (Wang et al., 2025) recently introduced a Multi-Agent Debate with Knowledge-Enhanced framework(MADKE). In their approach, agents access external evidence collected through a one-time, static retrieval process conducted prior to the debate. Although this method enhances factual grounding compared to purely internal MAD approaches, the fixed nature of the external knowledge prevents agents from adapting their evidence base as new arguments or knowledge gaps emerge during the debate.

In this paper, we introduce Tool-MAD, a dynamic MAD framework designed to enhance factual verification by equipping agents with external tools, including search APIs and Retrieval-Augmented Generation (RAG) modules (Lewis et al., 2020). Unlike previous MAD frameworks that rely on static or internally stored knowledge, Tool-MAD allows agents to dynamically and iteratively retrieve new evidence as the debate progresses. Specifically, after each debate round, agents formulate updated queries based on the arguments presented, enabling adaptive retrieval of relevant documents. This iterative knowledge collection significantly enhances the framework's adaptability, reduces hallucination risks, and improves the factual reliability of the final decision.

We conduct extensive experiments across four fact verification benchmark datasets to evaluate the effectiveness of Tool-MAD. The results show that Tool-MAD consistently outperforms competitive multi-agent debate framworks such as MAD (Liang et al., 2024) and MADKE (Wang et al., 2025), achieving performance improvements of up to 35.5 % and 8.0 %, respectively. Tool-MAD further demonstrates its flexibility in medical QA settings, maintaining robust performance under different retrieval tools and corpus configurations.

The main contributions of this paper can be summarized as follows:

- We propose Tool-MAD, a novel multi-agent debate framework enabling agents to verify factual claims dynamically using external tools such as search APIs and RAG modules.

- We introduce an adaptive query formulation mechanism, enabling agents to iteratively refine their evidence retrieval based on evolving debate contexts and previous arguments, leading to more informed and reliable judgments.

- We introduce the consistency score, a novel metric that quantifies the semantic similarity between an agent's response and its retrieved evidence, allowing the Judge agent to assess factual alignment and detect hallucinations.

- We comprehensively evaluate Tool-MAD on four benchmark datasets for fact verification, as well as two additional datasets for medical QA tasks, consistently surpassing competitive multi-agent debate baselines.

2 Related Works

Multi-Agent Debate Multi-agent debate frameworks, inspired by the "society of minds" theory (Minsky, 1986), aim to improve reasoning and decision-making through agent collaboration. For example, Du et al. (Du et al., 2023) introduced a multi-agent debate framework in which LLMs interact over multiple rounds, showing gains in both reasoning and factual consistency. More recently, Liang et al. (Liang et al., 2024) addressed the *"Degeneration of Thought"*, a phenomenon where self-reflection (Shinn et al., 2023b) fails to produce novel ideas, by introducing a "tit-for-tat" based multi-agent debate framework. Furthermore RECONCILE (Chen et al., 2023a) is a framework in which diverse LLM agents debate in a roundtable setting and reach a consensus answer via confidence weighted voting. Additionally, Wang et al. (Wang et al., 2025) introduced a knowledge-enhanced multi-agent debate setting, where agents select evidence from a shared document pool retrieved through external search engines. We note that unlike this existing MAD approach, the proposed Tool-MAD framework dynamically updates and expands its evidence pool throughout the debate, allowing agents to iteratively retrieve contextually relevant documents based on evolving arguments.

Fact Verification LLMs demonstrate outstanding natural language generation capabilities, but are often limited by their propensity to produce hallucinations-information that deviates from factual reality (Li et al., 2024; Chen et al., 2023b). As a result, research has increasingly focused on improving the factual consistency and trustworthiness of LLM-generated responses (Lee et al., 2020; Kim et al., 2024; Augenstein et al., 2024; Fadeeva et al., 2024). Building on prior research demonstrating that the integration of external knowledge sources, such as RAG (Lewis et al., 2020), can mitigate hallucinations and improve fact verification (Zhao et al., 2023), recent studies have proposed frameworks that combine LLMs with retrieval mechanisms to ground their outputs in reliable evidence and enhance factual consistency (Min et al., 2023).

Tool-Augmented Agents Recent advances in language model research have highlighted the ben-

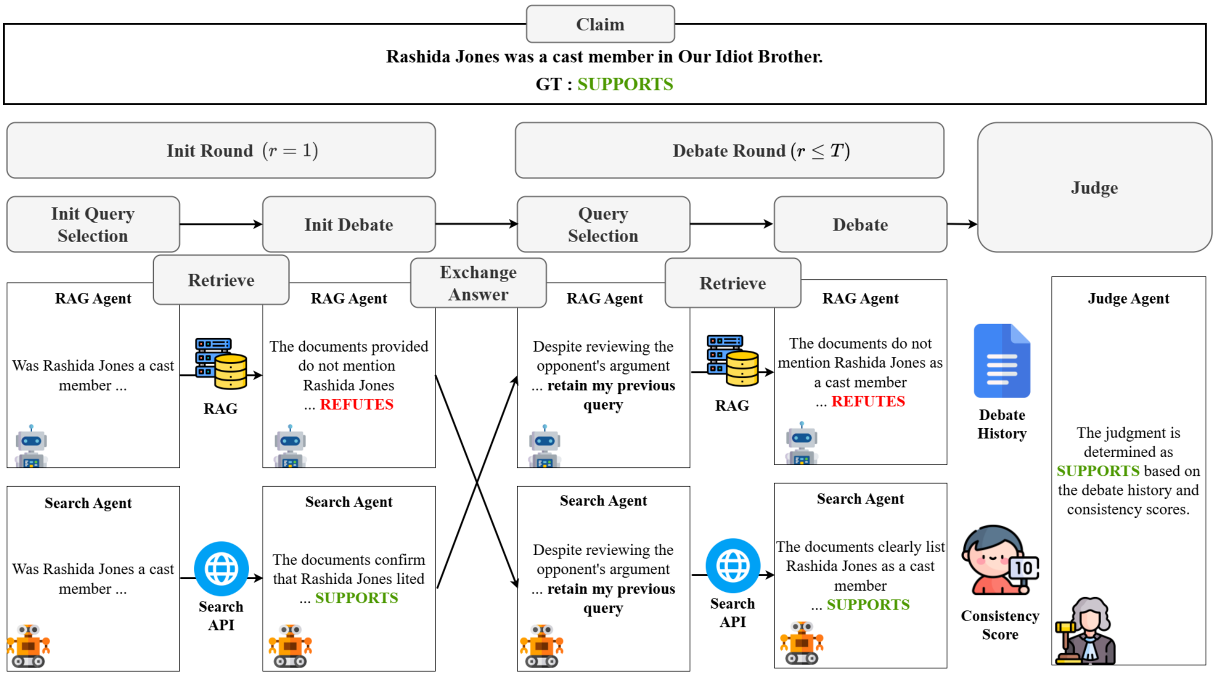

Figure 1: Given a claim, two agents (RAG and Search) engage in multi-round debates, where r denotes the current round and T is the predefined round threshold. If no consensus is reached by round T, the Judge Agent issues a final verdict based on debate history and consistency scores. Detailed examples are provided in the Appendix D.

efits of augmenting LLMs with external tools to overcome their inherent limitations and improve task performance (Schick et al., 2023; Shen et al., 2023; Lu et al., 2023; Surís et al., 2023). Furthermore, leveraging domain-specific tools for external interaction has been shown to enhance performance in specialized tasks (Bran et al., 2023; Gao et al., 2024).

3 Tool-MAD: Dynamic Multi-Agent Debate Framework

In this section, we introduce Tool-MAD, a novel MAD framework designed to enhance the factual reliability of LLMs for claim verification tasks. Unlike previous methods that rely primarily on static evidence or single-agent reasoning, Tool-MAD incorporates iterative retrieval of external evidence and dynamic interactions among multiple specialized agents.

3.1 External Tools

To enable dynamic, context-aware fact verification, Tool-MAD equips agents with external retrieval tools that provide access to relevant and timely evidence during debates. Specifically, we integrate two complementary retrieval mechanisms: a RAG module leveraging a static corpus, and a live web Search API for real-time information access.

Retrieval-Augmented Generation We employ the RAG framework (Guu et al., 2020) to augment agent reasoning with relevant documents retrieved from an embedded vector store. For this purpose, we use Milvus (Wang et al., 2021), a scalable and efficient vector database optimized for high-dimensional corpus management. We index a corpus constructed from Wikipedia articles, enabling rapid semantic retrieval. At inference time, each query returns the top three most semantically relevant documents, providing agents with targeted supporting evidence.

Search API Complementing the static RAG corpus, Tool-MAD also incorporates a real-time Search API, enabling agents to access up-to-date information directly from the web. For this, we utilize the Tavily Search API [1], known for its effective integration with language models. Similar to the RAG system, the Search API retrieves the three most relevant documents per query, ensuring comprehensive coverage of evolving knowledge demands during the debate.

3.2 Debate Participants

Debater Agents Tool-MAD involves two specialized debater agents: a RAG-based agent (A_R) that retrieves evidence from a static vector-based cor-

[1] https://tavily.com/

pus, and a Search-based agent (A_S) that accesses live web documents via a search API. Both agents operate under a shared prompt template to ensure consistent behavior across rounds. In the first round, each agent independently generates a response based solely on the input claim. In subsequent rounds, agents refine their responses by incorporating the previous response from their opponent, enabling richer reasoning and exposure to diverse viewpoints.

Judge Agent If the two debater agents fail to reach consensus within a predefined maximum number of rounds, the final decision is made by a third agent, the Judge (A_J). The Judge determines the outcome based on three primary inputs: (1) the original claim, (2) the full debate history—including agent responses and retrieved evidence from each round, and (3) a consistency score that assesses the factual alignment of each agent's arguments.

3.3 Tool-MAD Procedure

Tool-MAD proceeds in multiple rounds of interaction between two debater agents, each equipped with distinct external tools. Let c denote the input claim, and $r \in \{1, \dots, T\}$ be the current debate round, where T is the maximum allowed number of rounds. The framework involves three agents: a retrieval-based debater A_R, a search-based debater A_S, and a Judge agent A_J. Each debater generates arguments grounded in retrieved evidence, and if consensus is not reached, the Judge agent produces the final verdict based on the accumulated dialogue and supporting information. A high-level schematic is shown in Figure 1, and Complete algorithmic details are provided in Algorithm 1 in the Appendix.

Initialization Round ($r = 1$). In the first round, each debater independently constructs an initial query based solely on the input claim c, without reference to the opponent's argument. This query is submitted to the agent's designated retrieval tool, which returns a set of top-k relevant documents. The agent then composes its response by reasoning over the claim and the retrieved evidence. For agent $A \in \{A_R, A_S\}$, the process is formally defined as:

$$q_A^1 = Query(A, c) \quad (1)$$

$$D_A^1 = Retrieve(A, q_A^1) \quad (2)$$

$$a_A^1 = Respond(A, D_A^1, c) \quad (3)$$

where q_A^1 is the retrieval query, D_A^1 is the set of retrieved documents, and a_A^1 is the generated response for round 1.

Debate Rounds ($r > 1$). In each subsequent round, agents refine their queries and responses by incorporating the opponent's previous answer. This enables dynamic evidence updates and encourages more robust reasoning. Each round consists of three steps: query formulation, evidence retrieval, and response generation. Specifically:

$$q_A^r = Query(A, c, a_{\bar{A}}^{r-1}) \quad (4)$$

$$D_A^r = Retrieve(A, q_A^r) \quad (5)$$

$$a_A^r = Respond(A, D_A^r, c, a_{\bar{A}}^{r-1}) \quad (6)$$

where $a_{\bar{A}}^{r-1}$ is the previous response from the opposing agent $\bar{A}$.

If both agents produce the same response in any round, the debate terminates early with a consensus. Otherwise, the current round is appended to the debate history. If no consensus is reached after T rounds, the Judge agent A_J determines the final outcome.

Judge Decision (triggered only when no consensus is reached). The Judge agent evaluates three key components: (1) the original claim c, (2) the full debate history—including all responses and retrieved documents from each round, and (3) a consistency score that quantifies how well each agent's responses align with its retrieved evidence.

For each agent $A \in \{A_R, A_S\}$, the consistency score is defined as:

$$CS_A = \frac{1}{T} \sum_{r=1}^{T} similarity(D_A^r, a_A^r) \quad (7)$$

where $similarity(\cdot, \cdot)$ is a semantic similarity function implemented via a cross-encoder model[2]. The consistency score measures how well an agent's response aligns with the evidence it retrieved in each round, and is averaged over all rounds. This allows the Judge to assess the factual consistency of each agent's reasoning and to identify hallucinations or unsupported claims. The final verdict is based on these consistency scores in conjunction with the debate history and the original claim.

[2] https://huggingface.co/cross-encoder

Model	FEVER	FEVEROUS	FAVIQ	AVERITEC	Average
DeepseekR1	71.0	67.5	**77.0**	49.0	66.1
GPT-4o	69.5	54.5	68.0	43.5	58.9
GPT-4o-mini	62.0	37.0	56.0	33.0	47.0
+ CoT(Zero-shot)	66.5	31.0	67.0	33.5	49.5
+ Single Agent(ReAct)	62.0	24.0	66.0	24.0	44.0
+ MAD	71.0	36.5	68.0	36.0	52.9
+ MADKE	**72.0**	66.0	75.5	58.5	68.0
+ Tool-MAD	**72.0**	**72.0**	76.5	**64.5**	**71.3**
Llama-3.3-70B(Inst)	69.5	49.0	64.0	51.0	58.4
+ Tool-MAD	73.0	74.0	77.0	63.5	71.9

Table 1: Main results on four fact verification datasets (FEVER, FEVEROUS, FAVIQ, and AVeriTeC). Tool-MAD consistently improves performance across both proprietary models (GPT-4o, DeepSeekR1) and open-source backbones (LlaMA-3.3-70B), while outperforming other multi-agent debate frameworks. Average scores across datasets are also reported to provide a comprehensive comparison. **Bold** indicates the highest Exact Match score.

Model	MedQA	PubMedQA
MAD	58.0	22.5
MADKE	74.0	21.5
Tool-MAD	**79.0**	**31.0**

Table 2: This table compares multi-agent debate frameworks and demonstrates Tool-MAD's relative performance. Exact Match (EM) scores of different models on MedQA and PubMedQA. **Bold** indicates the highest score.

4 Experiments

We evaluate the effectiveness of a Tool-MAD framework that leverages external tools in performing the fact verification task. Unless specified otherwise, all experiments were conducted on 200 randomly sampled instances per dataset, evaluated using an Exact Match (EM) criterion, where a prediction is considered correct if it exactly matches the ground truth labels, and implemented with GPT-4o-mini (OpenAI, 2024) as the backbone model for Tool-MAD.

4.1 Datasets

Evaluation Datasets We evaluate our method across a wide range of tasks to assess both factual accuracy and cross-task flexibility. For fact verification, we utilize four widely used benchmark datasets: FEVER (Thorne et al., 2018) and FEVEROUS (Aly et al., 2021), which are based on Wikipedia-derived claims, as well as FAVIQ (Park et al., 2021) and AVERITEC (Schlichtkrull et al., 2023), which include claims from real world contexts. These datasets collectively span diverse domains and claim structures, allowing for a robust assessment of factual verification capabilities. To evaluate the flexibility of our framework beyond fact verification, we additionally include medical QA datasets such as MEDQA (Jin et al., 2021) and PubMedQA (Jin et al., 2019), which focus on domain-specific clinical and biomedical reasoning. Detailed descriptions of all datasets are provided in Appendix B.

4.2 Models

Backbone Model We employ GPT-4o-mini (OpenAI, 2024) and Llama-3.3-70B-Instruct-Turbo (Grattafiori et al., 2024) as backbone models in our Tool-MAD framework. We also utilize GPT-4o (OpenAI, 2024), a larger variant of GPT-4o-mini, for performance comparison. To further assess the effectiveness of the proposed framwork, we conduct a performance comparison with DeepseekR1 (Guo et al., 2025), a representative reasoning based model. More details are provided in Appendix C.1.

Baseline Model We evaluate Tool-MAD by comparing it with single agent reasoning and multi-agent debate baselines. See the Appendix E for a detailed structural comparison. A brief summary is provided below:

- **Zero-shot CoT** (Kojima et al., 2022) : Zero-shot CoT is to induce reasoning with the prompt "Let's think step by step".

- **Single Agent** (Yao et al., 2023) : The single-agent baseline is based on the ReAct framework, using RAG for external knowledge retrieval. ReAct enables iterative reasoning via tool use and feedback.

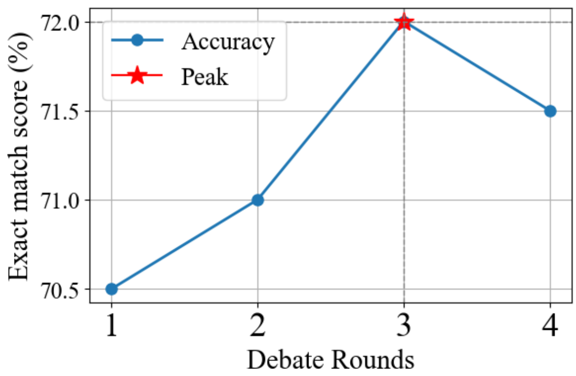

Figure 2: Exact Match performance of Tool-MAD on the FEVER dataset across different debate rounds.

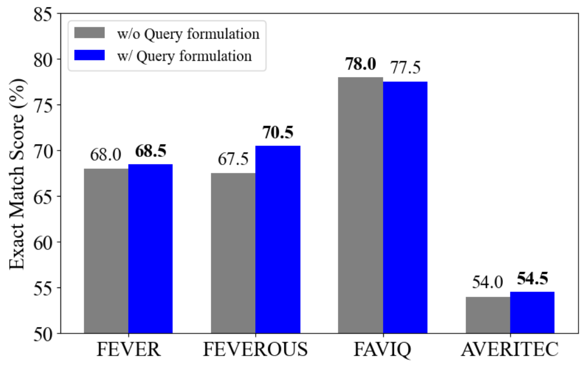

Figure 3: Exact Match performance comparison with and without query formulation across four benchmark datasets.

- **MAD** (Liang et al., 2024) : MAD is a framework that uses an interactive "tit for tat" debate structure, motivated by the Degeneration of Thought observed in single agents, even when self-reflection is applied.

- **MADKE** (Wang et al., 2025) :MADKE addresses the limitations of traditional MAD methods that rely solely on internal knowledge by incorporating a static evidence pool retrieved prior to the debate.

4.3 Fact Verification Experiments

We evaluate multi-agent debate models on four fact verification benchmark datasets. Table 1 presents the experimental results. Tool-MAD, using the lightweight GPT-4o-mini as its backbone, outperforms the more powerful GPT-4o across all evaluated benchmark datasets. Notably, it shows a 21% improvement on AVERITEC, the dataset with the highest label complexity, demonstrating the robustness of our framework under more challenging verification settings. Compared to DeepSeekR1, a prominent reasoning-based model, Tool-MAD achieves higher accuracy on three tasks, with gains of up to 15.5%. Furthermore, even with the open-source Llama-3.3-70B backbone, Tool-MAD surpasses both the standalone Llama-3.3-70B model and DeepSeekR1, highlighting its strong generalizability across different model backbones.

Tool-MAD outperforms other multi-agent debate frameworks, achieving up to 35.5% and 8.0% improvements over MAD and MADKE, respectively, with average gains of 18.4% and 3.3%. These results demonstrate that Tool-MAD delivers superior performance in fact verification tasks compared to existing multi-agent frameworks, and further highlight its architectural advantage over MADKE, which also leverages external knowledge. A detailed analysis of these results is provided in Section 5.1.

4.4 Flexibility

This experiment focuses on evaluating whether Tool-MAD can effectively adapt and maintain consistent performance when external tools are changed or applied to different domains. For comparison, we conduct flexibility experiments using the frameworks that showed competitive performance in Section 4.3, in place of single-agent baselines that demonstrated relatively lower performance. We evaluate our framework on two medical QA datasets: MedQA, which focuses on clinical multiple-choice questions, and PubMedQA, which targets biomedical fact verification. To test the framework's extensibility across tools, we configure the MedQA experiment using a PubMed-based RAG corpus, while the PubMedQA setting replaces the standard search API with OpenAlex (Priem et al., 2022), an open scholarly database.

In the PubMedQA pipeline, the agent extracts keywords from each query, retrieves three relevant abstracts, summarizes them using BERT (Devlin et al., 2019), and incorporates the summaries into the debate. Detailed information on the PubMedQA workflow is provided in the Appendix C.2.

As shown in Table 2, Tool-MAD outperforms other multi-agent debate frameworks on both datasets. Notably, while MADKE performs worse than MAD on PubMedQA despite incorporating external knowledge, Tool-MAD achieves strong results and maintains consistent performance even when the underlying tools are changed.

Model	Debater	FEVER	FEVEROUS	FAVIQ	Average
All VANILA	2	68.0	46.0	63.5	59.1
RAG + VANILA	2	70.0	55.0	69.5	64.8
SEARCH + VANILA	2	69.0	**68.5**	<u>74.0</u>	70.5
RAG + SEARCH + VANILA	3	**74.0**	65.5	73.0	<u>70.8</u>
Tool-MAD (RAG + SEARCH)	2	<u>72.5</u>	<u>66.0</u>	**74.5**	**71.0**

Table 3: Performance comparison of different agent combinations on FEVER, FEVEROUS, and FAVIQ datasets.Tool-MAD achieves the best overall results across combinations of RAG, SEARCH, and VANILA agents. **Bold** indicates the highest Exact Match score, while <u>underline</u> represents the second highest.

4.5 Ablation Study

Number of Debate Rounds We investigate how the number of debate rounds affects the overall performance of Tool-MAD. To analyze the effect of debate round, we conduct an experiment on the FEVER dataset. The results are shown in Figure 2. We observe that performance is lowest when only the Init Round (Round 1) is used. Accuracy gradually improves up to round 3 but slightly declines at round 4. Based on this observation, we set the threshold for the number of debate rounds to 3 in order to balance performance and efficiency.

Effect of Query Formulation We conducted ablation experiments to evaluate the effectiveness of dynamic query formulation. In Tool-MAD, each agent can independently revise its query at every debate round to retrieve more relevant documents. To isolate the impact of this mechanism, we compare the standard Tool-MAD setup with dynamic query updates at each round (w/ Query Formulation) to a variant that uses only the initial claim for retrieval throughout all rounds (w/o Query Formulation).

As shown in Figure 3, dynamic query formulation improves performance on most datasets, with the largest gain on FEVEROUS (+3.0%). While FaVIQ shows a slight drop, the overall trend confirms its positive contribution to fact verification accuracy. This highlights the importance of dynamically retrieving information based on the opponent's perspective, which enables access to more diverse evidence and leads to better performance.

Combination of Agents To evaluate the impact of different agent configurations, we compare their performance across the FEVER, FEVEROUS, and FAVIQ benchmark datasets. All agents follow the standard Tool-MAD procedure, with the only variation being the external tools they use. Specifically, we define three agent types: a base agent without external tools (VANILLA), an agent using retrieval-augmented generation (RAG), and an agent equipped with a web-based search API (SEARCH).

The results are presented in Table 3. As shown, configurations involving tool usage consistently outperform those without tools, demonstrating that external tool integration improves performance in fact verification tasks. Notably, the combination of RAG, SEARCH, and VANILLA performs slightly worse than the original Tool-MAD configuration (RAG + SEARCH), despite involving a larger number of debaters. The original Tool-MAD setup (RAG + SEARCH) achieved the highest average accuracy among all tested agent combinations, suggesting that it is a robust and effective configuration.

5 Discussion

5.1 Detailed Analysis on Fact Verification

This section analyzes the baseline models used in the fact verification experiments. Despite leveraging the ReAct framework and external knowledge, the Single Agent approach often produced incorrect conclusions, especially when retrieving irrelevant documents. MAD (Liang et al., 2024) alleviated some of these issues through debate, achieving better overall performance than the Single Agent, but still struggled due to its reliance on internal knowledge, often defaulting to "Not Enough Info" when evidence was insufficient.

MADKE, which incorporates external knowledge through a static evidence pool retrieved prior to the debate, demonstrated competitive performance against Tool-MAD. However, on more complex benchmarks such as AVERITEC and FEVEROUS, it showed a significant performance gap, highlighting the advantage of Tool-MAD's ability to dynamically access new information during the debate.

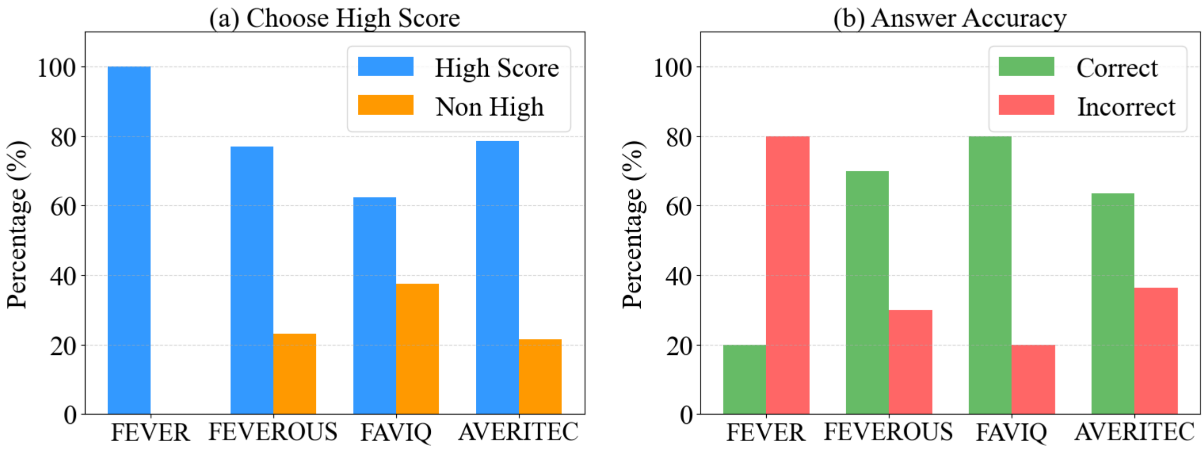

Figure 4: Evaluation of the role of consistency scores in LLM-based judgment. (a) Judges tended to prefer responses with higher consistency scores, selecting them in the majority of cases. (b) Higher consistency scores were strongly associated with correct responses, supporting the metric's validity as a factual accuracy.

5.2 Consistency Score

Figure 4 presents a summary of consistency score evaluation across four fact verification datasets, including FEVER, FEVEROUS, FAVIQ, and AVeriTeC. We investigate the utility of the consistency score, which serves as a semantic similarity measure between retrieved documents and agent-generated responses, in supporting factual judgment by large language models. Specifically, we evaluate (1) whether the Judge Agent tends to prefer responses with higher consistency scores, and (2) whether such responses are more likely to be factually correct.

As shown in Figure 4(a), responses with higher consistency scores tend to be preferred during the judgment process and also exhibit a higher alignment with the correct answers. Notably, in FEVER and AVeriTeC, high-consistency responses were chosen in 100% and approximately 78% of the cases, respectively, indicating that the judge strongly relies on the consistency score as a key decision-making metric.

Figure 4(b) shows that responses with higher consistency scores are generally more accurate across datasets. FAVIQ achieved the highest accuracy at 80%, followed by FEVEROUS and AVeriTeC, both exceeding 70%, indicating a strong correlation between consistency and correctness. In contrast, FEVER showed only 20% accuracy, mainly due to cases where neither agent's response matched the ground truth, often caused by insufficient or ambiguous retrieved evidence. Overall, the consistently high accuracy across datasets suggests that the consistency score is a promising and interpretable reliability metric in multi-agent decision-making.

6 Conclusion

In this paper, we introduce Tool-MAD, a dynamic multi-agent debate framework for fact verification that integrates external tools into the reasoning process. Each agent is equipped with distinct retrieval tools, such as RAG and real-time search APIs, allowing agents to gather relevant evidence during the debate, leading to more accurate claim verification The framework also introduces a consistency score to help the Judge agent assess semantic alignment between retrieved documents and agent responses, improving factual reliability and reducing hallucinations. Extensive experiments on multiple benchmarks show that Tool-MAD outperforms other multi-agent debate frameworks, demonstrating superior performance in both fact verification and generalization to medical QA tasks. These results position Tool-MAD as a promising and flexible foundation for future research in tool-augmented multi-agent reasoning.

7 Limitation

While Tool-MAD demonstrates strong performance across various tasks, it also has limitations.First its effectiveness heavily relies on the quality of external tools such as RAG and search APIs. Outdated or irrelevant information can reduce accuracy. Second, the framework introduces additional complexity and resource demands due to multi-agent coordination, tool integration, and consistency score computation, which may hinder scalability. Furthermore, early-stage errors from one agent can propagate through the debate, potentially leading to flawed final reasoning despite the use of a consistency score.

References

Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. Feverous: Fact extraction and verification over unstructured and structured information. *arXiv preprint arXiv:2106.05707*.

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, et al. 2024. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence*, 6(8):852–863.

Andres M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. 2023. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. 2023a. Reconcile: Round-table conference improves reasoning via consensus among diverse llms. *arXiv preprint arXiv:2309.13007*.

Yuyan Chen, Qiang Fu, Yichen Yuan, Zhihao Wen, Ge Fan, Dayiheng Liu, Dongmei Zhang, Zhixu Li, and Yanghua Xiao. 2023b. Hallucination detection: Robustly discerning reliable answers in large language models. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 245–255.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, et al. 2023. Palm-e: An embodied multimodal language model.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, et al. 2024. Fact-checking the output of large language models via token-level uncertainty quantification. *arXiv preprint arXiv:2403.04696*.

Shanghua Gao, Ada Fang, Yepeng Huang, Valentina Giunchiglia, Ayush Noori, Jonathan Richard Schwarz, Yasha Ektefaie, Jovana Kondic, and Marinka Zitnik. 2024. Empowering biomedical discovery with ai agents. *Cell*, 187(22):6125–6151.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.

Kyungha Kim, Sangyun Lee, Kung-Hsiang Huang, Hou Pong Chan, Manling Li, and Heng Ji. 2024. Can llms produce faithful explanations for fact-checking? towards faithful explainable fact-checking via multiagent debate. *arXiv preprint arXiv:2402.07401*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Nayeon Lee, Belinda Z Li, Sinong Wang, Wen-tau Yih, Hao Ma, and Madian Khabsa. 2020. Language models as fact checkers? *arXiv preprint arXiv:2006.04102*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. The dawn after the dark: An empirical study on factuality hallucination in large language models. *arXiv preprint arXiv:2401.03205*.

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. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904, Miami, Florida, USA. Association for Computational Linguistics.

Yining Lu, Haoping Yu, and Daniel Khashabi. 2023. Gear: Augmenting language models with generalizable and efficient tool resolution. *arXiv preprint arXiv:2307.08775*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*.

Marvin Minsky. 1986. *Society of mind*. Simon and Schuster.

OpenAI. 2024. Introducing gpt-4o: our fastest and most affordable flagship model. `https://platform.openai.com/docs/guides/vision`. Accessed: 2025-05-19.

Jungsoo Park, Sewon Min, Jaewoo Kang, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2021. Faviq: Fact verification from information-seeking questions. *arXiv preprint arXiv:2107.02153*.

Jason Priem, Heather Piwowar, and Richard Orr. 2022. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, volume 36, pages 68539–68551. Curran Associates, Inc.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. Averitec: A dataset for real-world claim verification with evidence from the web. *Advances in Neural Information Processing Systems*, 36:65128–65167.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36:38154–38180.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023a. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023b. Reflexion: language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pages 8634–8652. Curran Associates, Inc.

Dídac Surís, Sachit Menon, and Carl Vondrick. 2023. Vipergpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11888–11898.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

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

Haotian Wang, Xiyuan Du, Weijiang Yu, Qianglong Chen, Kun Zhu, Zheng Chu, Lian Yan, and Yi Guan. 2025. Learning to break: Knowledge-enhanced reasoning in multi-agent debate system. *Neurocomputing*, 618:129063.

Jianguo Wang, Xiaomeng Yi, Rentong Guo, Hai Jin, Peng Xu, Shengjun Li, Xiangyu Wang, Xiangzhou Guo, Chengming Li, Xiaohai Xu, et al. 2021. Milvus: A purpose-built vector data management system. In *Proceedings of the 2021 International Conference on Management of Data*, pages 2614–2627.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

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

Yiran Zhao, Jinghan Zhang, I Chern, Siyang Gao, Pengfei Liu, Junxian He, et al. 2023. Felm: Benchmarking factuality evaluation of large language models. *Advances in Neural Information Processing Systems*, 36:44502–44523.

A Prompts for Tool-MAD

We include three core prompt types used in the Tool-MAD framework: (1) Initial Round Prompts, which include query selection prompts used to generate initial search queries from the claim (Figure **??**), and answer generation prompts used by both RAG and Search agents to produce initial responses based solely on the retrieved documents and the original claim (Figure 6). (2) Debate Round Prompts, which incorporate the opponent's previous answer along with newly retrieved evidence to encourage cross-agent reasoning (Figure 7, Figure 8) and (3) Judge Prompts, used to synthesize the full debate history and consistency scores to render a final decision (Figure 9). Example templates for each type are provided below. The example in the figure is a FEVER prompt used by the RAG agent.

B Dataset details

To evaluate the performance of the proposed Tool-MAD framework across various task types, we utilized a total of eight publicly available datasets. These datasets are categorized into three task types, Fact Verification, Reasoning, and Medical. The datasets are summarized in the table 4.

FEVER is a dataset consisting of claims generated by modifying Wikipedia sentences, and each claim is labeled with one of three categories: Supported / Refuted / Not Enough Info.

FEVEROUS is a dataset that, unlike previous fact verification datasets which focused solely on unstructured text, incorporates both unstructured and structured information. Each claim is labeled as Supports, Refutes, or Not Enough Info.

FAVIQ is a challenging fact verification task constructed from claims that are likely to cause user confusion. Each claim is labeled as either Supports or Refutes.

AVERITEC is a dataset designed to overcome the limitations of exisiting fact-checking datasets (e.g., artificial claims, temporally inconsistent evidence). It consists of four labels, such as Supported, Refuted, Not Enough Evidence, and Conflicting Evidence/Cherrypicking, making it the dataset with the most diverse label set used in this study. In particular, the Conflicting Evidence/Cherrypicking label goes beyond traditional "Not Enough Evidence" cases by accounting for situations where evidence exists but is contradictory or selectively cited.

MEDQA & PUBMEDQA MedQA is a medical QA dataset based on multiple-choice questions from professional medical licensing exams. PubMedQA is a QA dataset constructed using abstracts from PubMed articles. MedQA requires selecting one correct option among multiple choices, while PubMedQA involves choosing one of three labels, such as Yes, No, Maybe.

C Experiments details

C.1 Using API

In this study, we utilized the OpenAI[3] for the GPT-4o and GPT-4o-mini models. For DeepSeekR1, we used the Novita API[4], and for the Llama-3.3-70B-Instruct Turbo model, we employed Together.ai API[5]. The temperature of all models was set to 0.0 to ensure reproducibility.

C.2 Flexibility

In the case of the reasoning task, the questions are descriptive and may have multiple ground truth answers. Therefore, an additional LLM module was incorporated into the original Tool-MAD framework to assess consensus. In the medical tasks, we explored the use of a new tool as a replacement the Search API. When keywords are selected, three relevant paper abstracts are retrieved, summarized using DistilBERT[6], and then provided to the agent. The workflow of this experiment is illustrated in Figure 10.

D Debate Examples

Final Round Judgement Example This section presents an example trajectory that illustrates the reasoning process leading to the final judgment (Figure 11). The example centers on the claim, "Is Rashida Jones a cast member of Our Idiot Brother?" Both the RAG and Search agents generate initial queries and perform retrieval. During this process, the RAG agent retrieves an incorrect document and argues that Rashida Jones is not part of the cast. In contrast, the Search agent accesses the correct document and argues that Rashida Jones is indeed listed as a cast member.

In the following round, both agents continue to stick to their own queries. The Search agent proceeds with the debate by citing a specific role

[3] https://openai.com/index/openai-api/
[4] https://novita.ai/
[5] https://www.together.ai/
[6] https://huggingface.co/sshleifer/distilbart-cnn-12-6

name, while the RAG agent persists in making an incorrect claim, arguing that the person does not appear in the cast list of the retrieved document. As a result, the two agents fail to reach an agreement by the final round. The Judge agent then makes the final decision based on the full debate history and the consistency scores. At this point, the Search agent has a higher consistency score and is shown to have reached the correct answer.

E Baseline Comparision

All models listed in Table 5 incorporate a reasoning component by design. Among them, only MAD, MADKE, and Tool-MAD adopt a multi-agent framework, enabling interaction between agents during the reasoning process. For external evidence access, retrieval capabilities are available in Single-Agent (ReAct), MADKE, and Tool-MAD. Furthermore, query formulation, which allows dynamic refinement of retrieval queries, is supported only by Single-Agent (ReAct) and Tool-MAD.

Notably, Tool-MAD is the only framework that combines all four structural components: reasoning, multi-agent interaction, retrieval, and dynamic query formulation into a single cohesive system. This integration allows Tool-MAD to not only access more relevant information during the debate but also dynamically adapt its evidence gathering strategy in response to opposing arguments, making it structurally more robust and flexible than other baseline methods.

You are a participant in a fact-checking debate.
Our goal is to reach a consensus with an accurate answer.
You are an agent utilizing RAG.
Choose an approrpiate query based on the given claim.\n\n
Only output the query, and wrap it in square brackets like this: [your query here].
Do not include anything else.\n\n

Claim:{prompt}\n

Figure 5: The prompt for initial round query selection

You are a participant in a fact-checking debate.
Our goal is to reach a consensus with an accurate answer.
You are an agent utilizing RAG.

If the claim is correct, output **reason** and **SUPPORTS**.\n
If the claim is incorrect, output **reason** and **REFUTES**\n
If the evidence is unclear or inconsistent, output **reason** and **NOT ENOUGH INFO**\n\n

Documents : {search_result}\n
Claim : {prompt}\n

Figure 6: The prompt for initial round debate

You are a participant in a fact-checking debate.
Our goal is to reach a consensus with an accurate answer. You are an agent utilizing RAG.\n\n

You are in query select page, you can choose change query or continue use your query\n\n
Only output the query, and wrap it in square brackets like this: [your query here].
Do not include anything else.\n\n

Other debaters answer {other_answer}\n
Before you used query : {used_query}\n
Claim : {prompt}\n

Figure 7: The prompt for query selection

You should determine your answer based on the documents you have retrieved and the other debaters answer, and the given claim.\n\n

If the claim is correct, output **reason** and **SUPPORTS**.\n
If the claim is incorrect, output **reason** and **REFUTES**\n
if the evidence is unclear or inconsistent, output **reason** and **NOT ENOUGH INFO**\n\n

Document : {search_result}\n
Other debaters answer : {other_answer}\n
Claim : {prompt}\n

Figure 8: The prompt for debate

```
You are the judge for the fact-checking debate.
This is the final round of the debate. If no consensus has been reached, you must determine the
correct answer.\n\n

The consistency score indicates how well the agent's statement aligns with the retrieved document\n
A higher score means better alignment and greater reliability\n
Based on the debate history and consistency score,
please determine the correctness of the claim as follows:\n\n

if the claim is correct, output **SUPPORTS**\n\n
if the claim is incorrect, output **REFUTES**\n\n
if the evidence is unclear or inconsistent, output  **NOT ENOUGH INFO**\n\n

Debate History: {debate_history}\n\n
RAG Agent's Consistency Score: {rag_consistency_score/3}\n\n
Search Agent's Consistency Score : {search_consistency_score/3}\n\n
Claim: {prompt}\n\n
```

Figure 9: The prompt for Judge

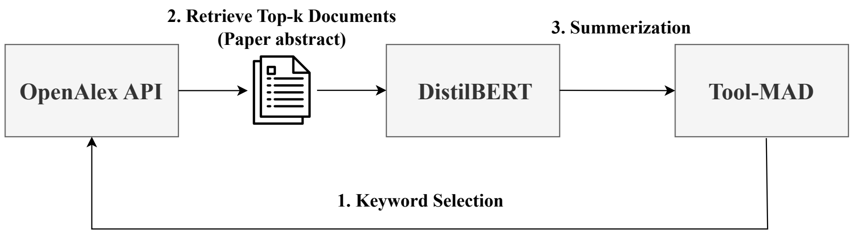

Figure 10: Workflow diagram of the paper retrieval API used in the PubMedQA experiment

Dataset	Reference	Task
FEVER	Thorne et al. (2018)	Fact Verification
FEVEROUS	Aly et al. (2021)	Fact Verification
FAVIQ	Park et al. (2021)	Fact Verification
AVERITEC	Schlichtkrull et al. (2023)	Fact Verification
MEDQA	Jin et al. (2021)	Medical
PUBMEDQA	Jin et al. (2019)	Medical

Table 4: Datasets used in Tool-MAD, categorized by task type and corresponding references.

Model	Reasoning	Multi-Agent	Retrieval	Query Formulation
CoT (Zero-shot)	O	X	X	X
Single-Agent (ReAct)	O	X	O	O
MAD	O	O	X	X
MADKE	O	O	O	X
Tool-MAD	O	O	O	O

Table 5: Structural comparison of baseline reasoning frameworks.

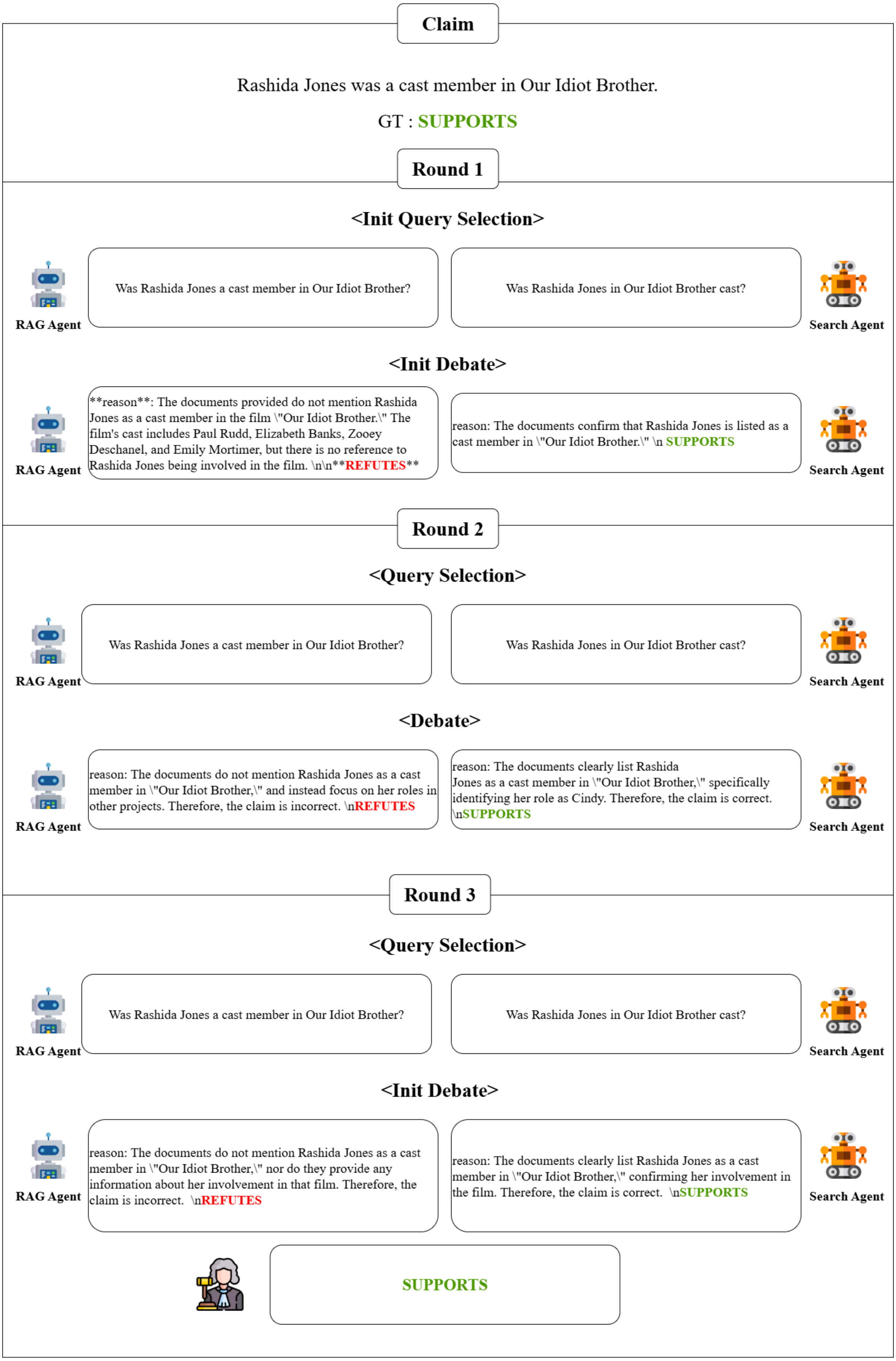

Figure 11: A detailed example of Figure 2

Algorithm 1 Tool-MAD Framework

Input: Claim c, Threshold Round T
Agents: RAG Agent A_R, Search Agent A_S, Judge Agent A_J
Initialize debate history $H \leftarrow \emptyset$,consistency scores $CS^R \leftarrow 0$, $CS^S \leftarrow 0$
for $r = 1$ to T **do** ▷ Iterate through debate rounds
  if $r = 1$ **then**
    Initial query based only on the claim:
    $q_R \leftarrow Query(A_R, c)$, $q_S \leftarrow Query(A_S, c)$
    Each agent retrieves documents and generates an answer:
    $d_R \leftarrow Retrieve\ (A_R, q_R)$, $a_R \leftarrow Respond(A_R, D_R, c)$
    $d_S \leftarrow Retrieve\ (A_S, q_S)$, $a_S \leftarrow Respond(A_S, D_S, c)$
  else
    Update query using opponent's last answer:
    $q_R \leftarrow Query(A_R, a_S, q_R, c)$, $q_S \leftarrow Query(A_S, a_R, q_S, c)$
    Each agent retrieves documents and generates an answer:
    $d_R \leftarrow Retrieve(A_R, q_R)$, $a_R \leftarrow Respond(A_R, D_R, c, a_S)$
    $d_S \leftarrow Retrieve(A_S, q_S)$, $a_S \leftarrow Respond(A_S, D_S, c, a_R)$
  end if
  if $a_R = a_S$ **then** ▷ Consensus reached
    return a_R
  end if
  Append round info to H ▷ Update Debate History
  $CS^R \leftarrow CS^R + similarity\ (a_R, D_R)$ ▷ Update Consistency Score
  $CS^S \leftarrow CS^S + similarity\ (a_S, D_S)$
end for
return Judge(A_J, H, CS^R, CS^S, c) ▷ Final decision by Judge agent