

Tool-MAD: A Multi-Agent Debate Framework for Fact Verification with Diverse Tool Augmentation and Adaptive Retrieval

Syeon Jeong¹, Yeonjun Choi¹, JongWook Kim², Beakcheol Jang¹

¹Graduate School of Information, Yonsei University, Seoul, Republic of Korea

²Department of Computer Science, Sangmyung University, Seoul, Republic of Korea

Abstract—Large Language Models (LLMs) suffer from hallucinations and factual inaccuracies, especially in complex reasoning and fact verification tasks. Multi-Agent Debate (MAD) systems aim to improve answer accuracy by enabling multiple LLM agents to engage in dialogue, promoting diverse reasoning and mutual verification. However, existing MAD frameworks primarily rely on internal knowledge or static documents, making them vulnerable to hallucinations. While MADKE introduces external evidence to mitigate this, its one-time retrieval mechanism limits adaptability to new arguments or emerging information during the debate. To address these limitations, We propose Tool-MAD, a multi-agent debate framework that enhances factual verification by assigning each agent a distinct external tool, such as a search API or RAG module. Tool-MAD introduces three key innovations: (1) a multi-agent debate framework where agents leverage heterogeneous external tools, encouraging diverse perspectives, (2) an adaptive query formulation mechanism that iteratively refines evidence retrieval based on the flow of the debate, and (3) the integration of Faithfulness and Answer Relevance scores into the final decision process, allowing the Judge agent to quantitatively assess the coherence and question alignment of each response and effectively detect hallucinations. Experimental results on four fact verification benchmarks demonstrate that Tool-MAD consistently outperforms state-of-the-art MAD frameworks, achieving up to 5.5% accuracy improvement. Furthermore, in medically specialized domains, Tool-MAD exhibits strong robustness and adaptability across various tool configurations and domain conditions, confirming its potential for broader real-world fact-checking applications.

Index Terms—LLM, Multi-Agent Debate, Fact Verification.

I. INTRODUCTION

LARGE Language Models (LLMs) have recently achieved strong performance across various NLP tasks [1]–[3], such as dialogue generation, summarization, and knowledge extraction. However, they often suffer from hallucination—producing confident yet factually incorrect content [4]. To mitigate this issue, prompt-based single-agent methods such as Chain-of-Thought (CoT) [5] and self-reflection [6] guide reasoning or incorporate external tools like the Wikipedia API. While simple and effective, these approaches often fail to revise incorrect answers, a limitation referred to as *Degeneration of Thought* [7].

Recently, multi-agent debate frameworks have been proposed to overcome this by enabling multiple LLMs to engage in argumentation and mutual verification [7], [8]. While promising, existing methods typically rely on static documents or internal model knowledge and still base final decisions

This work was supported by the National Research Foundation of Korea (NRF) funded by Korean Government under Grant RS-2023-00273751.

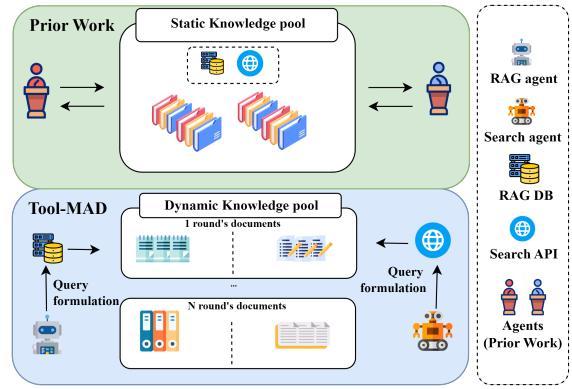


Fig. 1. Comparison of MAD [7], MADKE [9] and the proposed Tool-MAD: Unlike MAD and MADKE, which relies on a fixed document pool for retrieval, Tool-MAD dynamically retrieves external documents at each round through query formulation.

on LLM-generated debate history, which remains vulnerable to hallucination. To enhance factual grounding, MADKE [9] introduced static external retrieval before the debate. However, this evidence remains fixed during discussion, limiting the agents' ability to adapt to new claims or knowledge gaps that emerge throughout the debate.

A key challenge underlying these limitations is that fact verification in practical scenarios rarely depends on a single, uniform source of evidence. Different external tools provide distinct advantages: search APIs offer broad coverage and access to real-time information, whereas Retrieval-Augmented Generation (RAG) modules retrieve semantically aligned and context-rich documents from curated corpora such as Wikipedia. Tasks involving breaking news or fast-evolving events require up-to-date external information, while tasks involving historically stable knowledge demand high-precision retrieval. Relying on a single type of retrieval often leads to systematic blind spots. Moreover, prior debate frameworks typically assume that evidence retrieved before the discussion is sufficient for the entire process, overlooking the dynamic nature of argumentation. In human debates, participants routinely refine their claims and gather new evidence when confronted with counterarguments or inconsistencies. Without a mechanism for iterative retrieval, debate agents remain restricted by the limitations of their initial evidence set.

In this paper, we introduce Tool-MAD, a dynamic multi-agent debate framework designed to enhance factual verification by enabling agents to leverage heterogeneous ex-

ternal tools such as search APIs and Retrieval-Augmented Generation (RAG) modules, thereby promoting diverse perspectives [10]. Unlike prior multi-agent debate frameworks that rely on static retrieval or internally stored knowledge within the model, Tool-MAD is designed to allow agents to dynamically and iteratively retrieve new evidence as the debate progresses. Specifically, after each debate round, agents update their queries based on previous arguments exchanged during the discussion, enabling adaptive retrieval of relevant documents. This iterative knowledge acquisition process significantly enhances the adaptability of the framework, reduces the likelihood of hallucinations, and contributes to improving the factual reliability of the final decision. This mechanism is illustrated in Figure 1, which compares MAD, MADKE and Tool-MAD.

In addition, we evaluate the responses generated in each debate round using two independent metrics from the RAGAS framework [11]: faithfulness, which measures how well an agent’s claim is grounded in the retrieved evidence, and answer relevance, which assesses how directly the response addresses the original question. As these two metrics capture complementary aspects of response quality, we collectively refer to them as the stability score, which serves as an auxiliary signal for the Judge agent to assess the factual consistency and trustworthiness of each response.

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 frameworks such as MAD [7] and MADKE [9], achieving performance improvements of up to 35.5 % and 5.5 %, 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 that empowers agents to verify factual claims adaptively by leveraging a diverse set of external tools, including real-time search APIs and Retrieval-Augmented Generation (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 incorporate faithfulness and answer relevance as a stability score in Tool-MAD to assess how well responses align with evidence and address the original question. This helps detect hallucinations and guide final decisions.
- 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.

II. RELATED WORKS

Multi-Agent Debate. Multi-agent debate frameworks have emerged as a promising direction for strengthening LLM rea-

soning by leveraging adversarial or collaborative interactions between agents. Early theoretical foundations stem from Minsky’s “society of minds” theory [12], which views intelligence as the emergent result of multiple interacting subsystems. This idea has resurfaced in modern LLM settings, where multiple agents exchange arguments, challenge assumptions, and refine intermediate reasoning.

Du et al. [8] demonstrated that multi-round debate enables LLMs to correct each other’s errors and improve logical consistency. Liang et al. [7] introduced a tit-for-tat debate structure to combat the *Degeneration of Thought* problem identified in self-reflection methods [13]. Other works have explored more complex multi-agent coordination strategies, such as majority voting, argument diversification, or structured deliberation trees. RECONCILE [14] positions agents as participants in a roundtable discussion, integrating confidence-weighted consensus to avoid dominance by a single agent.

Beyond general reasoning, multi-agent frameworks have also been applied in specialized settings such as planning, safety evaluation, and code generation. For instance, debate-based self-correction mechanisms have been explored in mathematical reasoning, where agents critique intermediate steps, and in safety-alignment contexts, where disagreement is used to uncover unsafe model [8]. These efforts show that debate can create richer reasoning traces, but they still rely heavily on internal knowledge.

MADKE [9] introduced the idea of knowledge-enhanced debate by injecting externally retrieved documents before discussion; however, the evidence pool remains unchanged throughout the debate process. Similarly, debate-enhanced RAG systems retrieve supporting evidence prior to deliberation and then freeze the document set. While these systems show empirical gains, the fixed-evidence assumption limits adaptability when new arguments emerge during the debate.

Fact Verification and Factuality Evaluation. Ensuring the factual correctness of LLM outputs is a long-standing challenge [4], [15]. Early fact verification approaches relied on supervised classification models over structured evidence sources such as Wikipedia or news corpora. With the rise of generative models, research has shifted toward grounding generation in external corpora through retrieval-augmented generation (RAG) [10]. RAG-based fact-checkers [16] and hybrid evidence selection frameworks have shown that integrating retrieval significantly reduces hallucination in open-domain QA and claim verification.

Fact verification evaluation has similarly diversified. Works such as FEVER and its successors introduced label-based evaluation pipelines. More recent frameworks such as FactScore [17] attempt to quantify factuality at the claim level by decomposing model outputs. Consistency-based factuality approaches, including self-consistency [18] and re-asking [19], evaluate factual correctness by measuring the stability of model outputs rather than relying solely on external grounding.

RAGAS [11] bridged retrieval-based reasoning and factuality evaluation by introducing two complementary metrics: *faithfulness*, measuring whether claims match retrieved evidence, and *answer relevance*, measuring whether responses address the question directly. Several subsequent works have

adopted faithfulness for hallucination detection in RAG systems or used relevance-based filtering to identify unusable generations. However, these metrics have predominantly been applied as post-hoc evaluators for fully generated outputs, not as real-time signals that influence multi-step reasoning.

Furthermore, no prior work integrates metric-guided evaluation into a multi-agent debate structure, nor uses factuality signals to modulate argument selection, judge decisions, or evidence refinement across rounds. Tool-MAD incorporates these signals internally, using faithfulness and answer relevance as round-level stability indicators, which is distinct from existing factuality scoring frameworks.

Tool-Augmented and Retrieval-Augmented Agents. A growing line of research explores augmenting LLMs with external tools, enabling them to perform tasks requiring specialized knowledge or computation. Toolformer [20] enables models to learn API-calling behaviors, while Hugging-GPT [21] and GEAR [22] treat the LLM as a coordinator for heterogeneous models or systems. Tool-use frameworks have since evolved into modular agent pipelines that combine information extraction, symbolic reasoning, multi-step planning, or domain-specific simulators.

In scientific and professional domains, tool-augmented agents such as ChemCrow [23] and biomedical retrieval agents [24] demonstrate that domain-specific tools can dramatically improve reasoning fidelity. Retrieval modules—including dense retrievers, hybrid rankers, and cross-encoder rerankers—enable models to ground their reasoning in curated evidence.

However, existing tool-augmented systems overwhelmingly adopt a single-agent perspective and use tools in a one-shot or sequential fashion. They do not exploit multi-agent interaction as a mechanism for tool selection, evidence diversification, or iterative grounding. Likewise, retrieval systems typically perform a single retrieval step at the beginning of a task, without adapting to evolving reasoning trajectories or emerging counterarguments. As a result, these systems often fail to revisit or refine earlier tool calls when new uncertainties arise, leading to brittle reasoning paths. Moreover, the lack of interaction between agents prevents the system from leveraging disagreement or complementary viewpoints to guide more targeted retrieval or tool use.

III. TOOL-MAD

In this section, we introduce Tool-MAD, a novel multi-agent debate 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. By repeatedly updating evidence and challenging one another’s conclusions, the system progressively refines its reasoning, mitigates hallucinations, and promotes more reliable consensus formation.

A. 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 [25] to augment agent reasoning with relevant documents retrieved from an embedded vector store. For this purpose, we use Milvus [26], 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 API¹, 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.

Algorithm 1 Tool-MAD Framework

```

Input: claim  $c$ , debater set  $\mathcal{A}$ , max rounds  $T$ , stability thresholds  $St = (f, ar)$ 
Output: final decision answer
Initialize history  $H \leftarrow \emptyset$ 
for each  $A \in \mathcal{A}$  do
     $S^A \leftarrow 0$ 
end for
for  $r = 1$  to  $T$  do
    for each agent  $A \in \mathcal{A}$  do
        Let  $\bar{A}$  be the opponent of  $A$ 
        if  $r = 1$  then
             $q_A^{(r)} \leftarrow Query(A, c)$ 
             $D_A^{(r)} \leftarrow Retrieve(A, q_A^{(r)})$ 
             $a_A^{(r)} \leftarrow Respond(A, D_A^{(r)}, c)$ 
        else
             $q_A^{(r)} \leftarrow Query(A, c, a_{\bar{A}}^{(r-1)}, q_A^{(r-1)})$ 
             $D_A^{(r)} \leftarrow Retrieve(A, q_A^{(r)})$ 
             $a_A^{(r)} \leftarrow Respond(A, D_A^{(r)}, c, a_{\bar{A}}^{(r-1)})$ 
        end if
         $f_A^{(r)} \leftarrow faithfulness(a_A^{(r)}, D_A^{(r)})$ 
         $ar_A^{(r)} \leftarrow answerRelevance(c, a_A^{(r)})$ 
         $S^A \leftarrow S^A + (f_A^{(r)}, ar_A^{(r)})$ 
    end for
    Append round- $r$  info to  $H$ 
    if  $a_R = a_S$  and  $S^R > St$  and  $S^S > St$  then
        return  $a_R$ 
    end if
end for
return  $Judge(A_J, H, \{S^A\}_{A \in \mathcal{A}}, c)$ 

```

¹<https://tavily.com/>

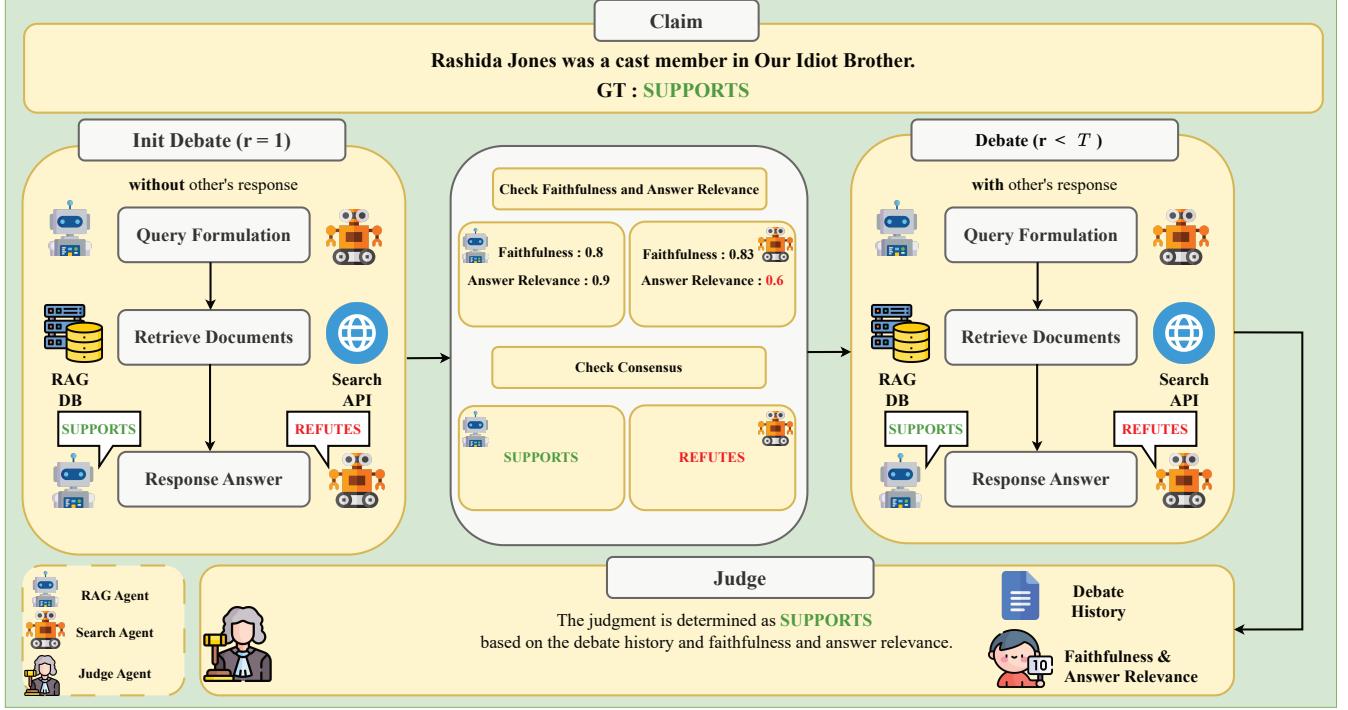


Fig. 2. Given a claim, two agents (RAG and Search) engage in multi-round debates, where r denotes the current round and T is the predefined round threshold. If no consensus is reached or the stability score falls below the threshold by round T , the Judge Agent issues a final verdict based on the debate history and the stability score.

B. 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 corpus, 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 stability score that assesses the factual consistency and relevance of each agent’s arguments.

C. Stability Score

To quantitatively evaluate the reliability of each agent’s response, Tool-MAD adopts two core metrics from the RAGAS framework [11]: *faithfulness* and *answer relevance*. These two metrics jointly form the *Stability Score* used throughout the debate.

Faithfulness measures how accurately an agent’s response reflects the content of the retrieved evidence. Each response

$a_s(q)$ is decomposed into a set of factual statements:

$$S(a_s(q)) = \{s_1, s_2, \dots, s_{|S|}\}.$$

For each statement s_i , the LLM determines whether it can be inferred from the retrieved context $c(q)$. We define a verification function $v(s_i, c(q))$ such that $v(s_i, c(q)) = 1$ if s_i is supported by $c(q)$, and $v(s_i, c(q)) = 0$ otherwise. The faithfulness score F is then computed as the proportion of supported statements:

$$F = \frac{\sum_{i=1}^{|S|} v(s_i, c(q))}{|S|}.$$

A high faithfulness score indicates that the response is well-grounded and factually consistent with the retrieved evidence.

Answer relevance assesses how directly the agent’s response addresses the original question. For the given answer $a_s(q)$, the LLM generates a set of n potential questions:

$$Q' = \{q_1, q_2, \dots, q_n\},$$

where each q_i is intended to represent a question that the answer could plausibly correspond to. We compute embeddings for the original question q and each generated question q_i , and calculate their cosine similarity:

$$\text{sim}(q, q_i) = \cos(\mathbf{e}(q), \mathbf{e}(q_i)).$$

The answer relevance score is then obtained by averaging the similarities:

$$AR = \frac{1}{n} \sum_{i=1}^n \text{sim}(q, q_i).$$

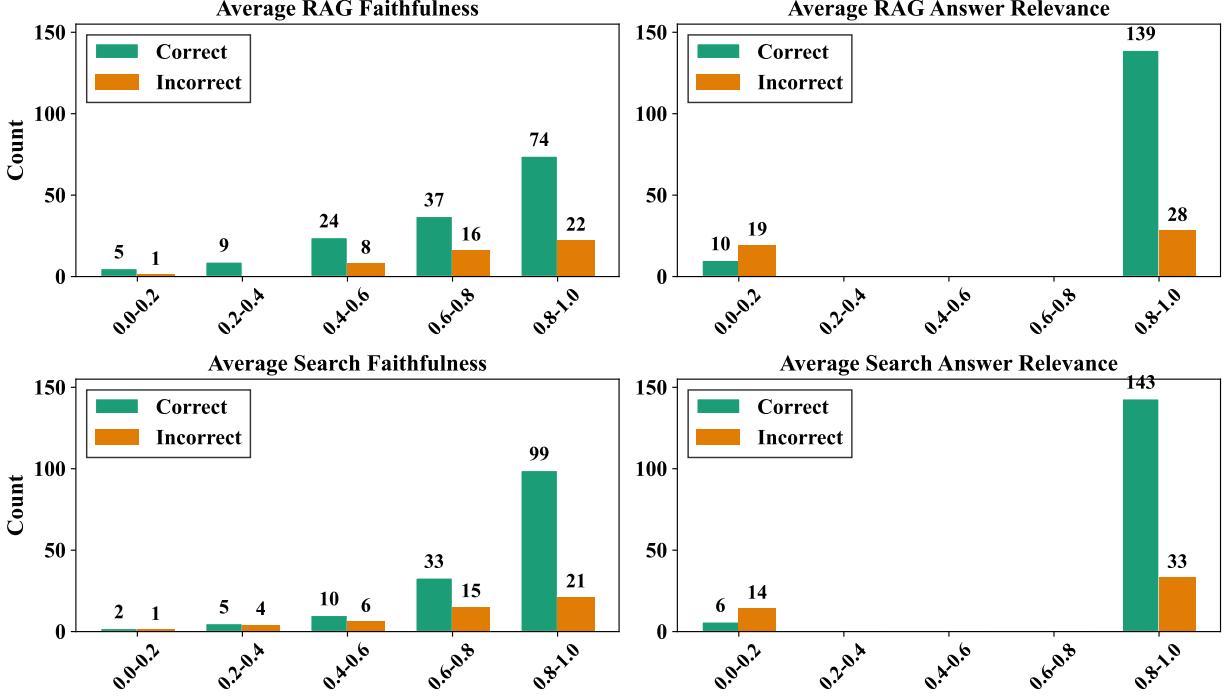


Fig. 3. Empirical distributions of faithfulness and answer relevance for selecting stability-score thresholds. Answer relevance is concentrated above 0.8, while faithfulness shows a wider spread around 0.7–0.8. We therefore set thresholds to 0.7 (faithfulness) and 0.8 (answer relevance) to balance precision and efficiency

A high relevance score suggests that the answer remains focused on the intended question, whereas a low score indicates incompleteness or topic drift.

Both metrics are computed at every debate round. If either faithfulness or answer relevance falls below a predefined threshold, the round is marked as inconclusive and the debate continues.

Threshold of Stability Score To determine suitable stability-score thresholds, we analyzed the empirical distributions of faithfulness and answer relevance (Figure 3). Answer relevance was strongly skewed toward high values, with most scores above 0.8, making 0.8 a natural cutoff that filters low-quality outputs without affecting the majority of valid ones. In contrast, faithfulness exhibited a wider spread, with many responses falling between 0.7 and 0.8; thus, using a stricter threshold would unnecessarily reject reasonable, evidence-aligned answers and prolong debates.

These observations reveal a key trade-off: overly strict thresholds increase precision but hurt efficiency, especially when models produce borderline yet acceptable outputs. Therefore, we adopt 0.7 for faithfulness and 0.8 for answer relevance, which empirically provides stable convergence while maintaining decision quality.

D. 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

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.

You are in query select page, you can choose change query or continue use your query.

Only output the query, and wrap it in square brackets like this: [your query here]. Do not include anything else.

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

Fig. 4. Prompt of query selection

retrieval-based debater A_R , a search-based debater A_S , and a Judge agent A_J . Each debater generates arguments based on its retrieved evidence, and the responses in each round are quantitatively evaluated using faithfulness and answer relevance scores. If either agent's response fails to meet predefined thresholds for these scores, or if no consensus is reached, the debate proceeds to the next round. If consensus is not achieved by the final round T , the Judge agent produces a final verdict based on the accumulated dialogue and retrieved evidence. The overall framework overview is presented in Figure 2, and the complete algorithmic details are provided in the algorithm 1.

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.

N Round's Debate
If the claim is correct, you must first explain why it is correct based on the document, then output **SUPPORTS**.
If the claim is incorrect, you must first explain why it is incorrect based on the document, then output **REFUTES**.
If the evidence is unclear or inconsistent, you must first explain the uncertainty, then output **NOT ENOUGH INFO**.
Your final answer (**SUPPORTS**, **REFUTES**, or **NOT ENOUGH INFO**) must appear on the last line only, after your reasoning.
Claim : {prompt} Other debaters answer : {other_answer} Document : {search_result}

Fig. 5. Prompt of round debate

Judge
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.
The faithfulness score measures how accurately the agent's answer reflects the retrieved documents, while the answer relevancy score indicates how well the answer addresses the original question.
A higher score means better alignment and greater reliability Based on the debate history and faithfulness and answer relevancy score, please determine the correctness of the claim as follows: if the claim is correct, output **SUPPORTS** if the claim is incorrect, output **REFUTES** if it is uncertain whether the claim is correct, output **NOT ENOUGH INFO**
Debate History : {debate_history} RAG Agent's Faithfulness Score : {rag_faithfulness / 3} RAG Agent's Relevancy Score : {rag_answer_relevance / 3} Search Agent's Faithfulness Score : {search_faithfulness / 3} Search Agent's Relevancy Score : {search_answer_relevance / 3} Claim : {prompt}

Fig. 6. Prompt of judge

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:

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

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

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

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:

TABLE I
DATASETS USED IN TOOL-MAD, CATEGORIZED BY TASK TYPE AND CORRESPONDING REFERENCES.

Dataset	Task
FEVER [27]	Fact Verification
FEVEROUS [28]	Fact Verification
FAVIQ [29]	Fact Verification
AVERITEC [30]	Fact Verification
MEDQA [31]	Medical
PUBMEDQA [32]	Medical

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

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

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

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, considering predefined thresholds of the stability score. Otherwise, the current round is appended to the debate history. If consensus is not reached or the responses fall below the thresholds for faithfulness or answer relevance, the debate continues until round T . The Judge agent A_J then determines the final outcome.

Judge Decision(triggered only when the debate reaches the final round) If the agents fail to reach a consensus by the final round T , the Judge agent A_J makes the final decision based on the original claim c , the complete debate history H , and the stability score S . Here, S is composed of the average faithfulness and answer relevance scores across all rounds, capturing the overall quality of the agents' responses. This decision process is formalized as:

$$\text{final answer} = \text{Judge}(A_J, H, S, c) \quad (7)$$

The LLM-based Judge agent receives the stability score S and debate history H as part of its prompt and makes the final decision by jointly considering both.

For reproducibility, we provide the full prompts used for the debate round, query selection, and judge components in Figures 4, 5, and 6. Note that the prompt for the initial round corresponds to the same template with the opponent's response removed.

IV. 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.

TABLE II
STRUCTURAL COMPARISON OF BASELINE REASONING FRAMEWORKS.

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 III

MAIN RESULTS ON FOUR FACT VERIFICATION DATASETS (FEVER, FEVEROUS, FAVI-Q, AND AVERITEC). TOOL-MAD CONSISTENTLY IMPROVES PERFORMANCE ACROSS BOTH PROPRIETARY (E.G., GPT-4O, DEEPSEEKR1) AND OPEN-SOURCE (E.G., LLAMA-3.3-70B) MODELS, OUTPERFORMING OTHER MULTI-AGENT DEBATE FRAMEWORKS. AVERAGE SCORES ARE REPORTED FOR OVERALL COMPARISON. **BOLD** INDICATES THE HIGHEST EXACT MATCH SCORE WITHIN EACH BASE MODEL GROUP (ABOVE: GPT-4 VARIANTS AND DEEPSEEKR1; BELOW: LLAMA).

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	73.0	71.5	77.5	62.0	71.0
LLama-3.3-70B(Inst)	69.5	49.0	64.0	51.0	58.4
+ CoT(Zero-shot)	71.0	51.0	68.5	32.0	55.6
+ Single Agent(ReAct)	73.5	51.0	65.5	35.5	56.4
+ MAD	54.0	31.5	58.5	39.5	45.9
+ MADKE	62.5	61.0	71.5	31.0	56.5
+ Tool-MAD	74.0	77.0	78.5	66.5	74.0

A. 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 [27] and FEVEROUS [28], which are based on Wikipedia-derived claims, as well as FAVI-Q [29] and AVERITEC [30], 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 [31] and PubMedQA [32], which focus on domain-specific clinical and biomedical reasoning. A detailed summary of all datasets, including scale, domain, and task type, is provided in Table I.

B. Models

We employ GPT-4o-mini [33] and Llama-3.3-70B-Instruct-Turbo [34] as backbone models in our Tool-MAD framework. We also utilize GPT-4o [33], a larger variant of GPT-4o-mini, for performance comparison. To further assess the effectiveness of the proposed framework, we conduct a performance comparison with DeepseekR1 [35], a representative reasoning based model.

Baseline Model We evaluate Tool-MAD by comparing it with single agent reasoning and multi-agent debate baselines. A brief summary is provided below:

- **Zero-shot CoT** [36] : Zero-shot CoT is to induce reasoning with the prompt “Let’s think step by step”.
- **Single Agent** [37] : The single-agent baseline is based on the ReAct framework, using RAG for external knowledge

TABLE IV
ACCURACY AND 95% CONFIDENCE INTERVALS ACROSS DATASETS.
(BACKBONE MODEL : GPT-4O-MINI)

Dataset	Acc.	95% C.I
FEVER	0.73	(0.67, 0.79)
FEVEROUS	0.72	(0.66, 0.78)
FAVIQ	0.78	(0.72, 0.83)
AVeriTec	0.62	(0.56, 0.69)

retrieval. ReAct enables iterative reasoning via tool use and feedback.

- **MAD** [7] : 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** [9] : 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.

Table II provides a structural comparison of these baselines, highlighting their differences across four key dimensions: intrinsic reasoning capability, multi-agent interaction, retrieval usage, and query formulation. As shown in the table, Tool-MAD is the only framework that simultaneously supports all four components by integrating structured debate, iterative retrieval, and adaptive query rewriting. In contrast, prior baselines typically cover only a subset of these capabilities, resulting in more limited overall functionality.

C. Fact Verification Experiments

We evaluate multi-agent debate models on four fact verification benchmark datasets. Table III presents the experimental

TABLE V
EXACT MATCH (EM) SCORES OF DIFFERENT MULTI-AGENT DEBATE FRAMEWORKS ON MEDQA AND PUBMEDQA. **BOLD** INDICATES THE HIGHEST SCORE.

Model	MedQA	PubMedQA
MAD	58.0	22.5
MADKE	74.0	21.5
Tool-MAD	77.0	29.0

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 18.5% improvement on AVERITEC, the dataset with the highest label complexity, demonstrating the robustness of our framework under more challenging verification settings. Compared to DeepSeekR1, Tool-MAD consistently achieves higher accuracy across all tasks, with improvements of up to 13%. Tool-MAD achieves comparable performance with the open-source Llama-3.3-70B backbone, confirming its robustness across model backbones.

Tool-MAD outperforms other multi-agent debate frameworks, achieving up to 35.0% and 5.5% improvements over MAD and MADKE, respectively, with average gains of 18.1% and 3.0%. 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.

To better understand the observed performance gap, we analyze the limitations of each baseline system. The Single Agent, despite leveraging the ReAct framework [37] and external tools, frequently generates incorrect answers due to the retrieval of irrelevant or incomplete documents. MAD [7] improves upon this through multi-agent debate, but still struggles when evidence is insufficient, often defaulting to a generic “Not Enough Info” response. MADKE [9] achieves relatively stable performance by relying on a static evidence pool retrieved prior to the debate, yet fails to generalize well on more complex datasets such as AVERITEC and FEVEROUS, where dynamic adaptation is required. These results highlight the architectural advantage of Tool-MAD, which enables agents to iteratively retrieve new and contextually relevant evidence throughout the debate, leading to improved factual reasoning.

Additionally, we report 95% bootstrap confidence intervals for Tool-MAD using GPT-4o-mini as the backbone. Due to the computational cost of multi-round multi-agent debate—where each query triggers multiple LLM inferences—conducting large-scale repeated trials is prohibitively expensive. The reported confidence intervals (Table II) therefore serve to supplement the robustness of our findings by quantifying the variability of the results under resampling. The corresponding results are presented in Table IV.

D. 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 Fact Verification Experiments section, 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 [38], an open scholarly database.

In the PubMedQA pipeline, the agent extracts keywords from each query, retrieves three relevant abstracts, summarizes them using BERT [39], and incorporates the summaries into the debate. Detailed information on the PubMedQA workflow is provided in the figure 7.

As shown in Table V, 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.

Beyond these quantitative results, the strong performance of Tool-MAD on both MEDQA and PubMedQA can be attributed to the complementary nature of its heterogeneous retrieval sources. Clinical QA tasks often require both up-to-date evidence, such as newly published studies or evolving treatment guidelines, and structured biomedical knowledge grounded in established definitions or mechanistic explanations. Search-based retrieval is well suited for capturing the former, whereas RAG-based retrieval excels at providing the latter. By integrating both types of information within the debate, Tool-MAD enables agents to form more comprehensive and reliable arguments compared to methods that rely on a single evidence modality.

Another noteworthy aspect is that Tool-MAD maintains stable performance despite the domain shift from Wikipedia-style fact verification to biomedical question answering. This robustness arises from the framework’s iterative retrieval mechanism, which allows agents to refine their queries and adjust their evidence sources as domain-specific reasoning challenges emerge. The ability to dynamically cross-validate retrieved evidence across rounds helps mitigate errors introduced by unfamiliar terminology or dataset-specific retrieval noise.

Finally, the debate structure used in Tool-MAD naturally aligns with multi-source reasoning patterns observed in clinical decision-support workflows. In practice, healthcare professionals often synthesize information from both standardized guidelines and recent studies, and Tool-MAD mirrors this process by enabling different agents to contribute distinct yet complementary evidence. The use of faithfulness and answer relevance as stability indicators further helps filter out unsupported or clinically unreliable responses, providing an additional layer of safety when operating in high-stakes medical domains.

TABLE VI

PERFORMANCE COMPARISON OF DIFFERENT AGENT COMBINATIONS ON FEVER, FEVEROUS, FAVIQ AND AVERITEC DATASETS. TOOL-MAD ACHIEVES THE BEST OVERALL RESULTS ACROSS COMBINATIONS OF RAG, SEARCH, AND VANILA AGENTS. **BOLD** INDICATES THE HIGHEST EXACT MATCH SCORE.

Model	FEVER	FEVEROUS	FAVIQ	AVERITEC
VANILA + VANILA	62.0	40.0	60.0	45.0
RAG + VANILA	65.5	57.5	63.5	58.0
SEARCH + VANILA	60.5	43.0	63.5	53.5
RAG + RAG	67.5	60.5	68.5	56.5
SEARCH + SEARCH	67.0	60.5	68.5	55.5
Tool-MAD (RAG + SEARCH)	73.0	71.5	77.5	62.0

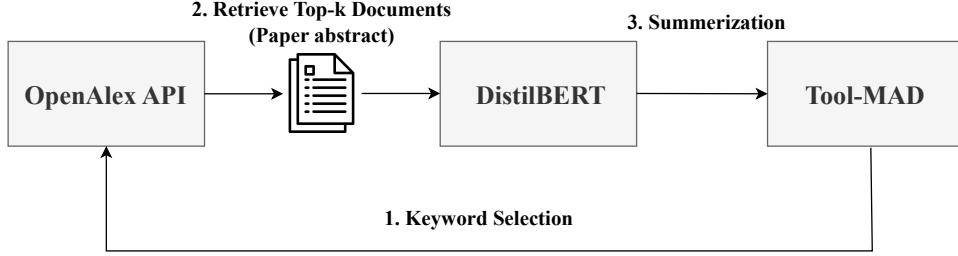


Fig. 7. Workflow diagram of the paper retrieval API used in the PubMedQA experiment

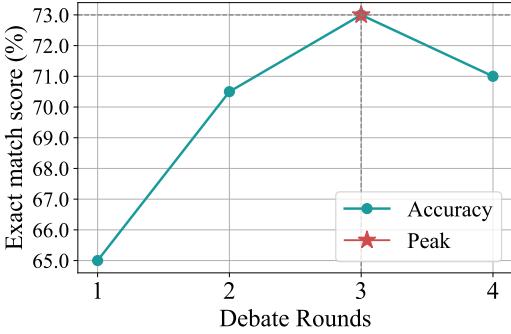


Fig. 8. Exact Match performance of Tool-MAD on the FEVER dataset across different debate rounds.

E. Ablation Study

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 same Tool-MAD procedure, with the only variation being the external tools they employ. Specifically, we define three types of agents: a base agent without any 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 VI. As shown in the table, configurations that incorporate external tools consistently outperform those without tools, indicating that tool integration improves performance in fact verification tasks. Among all agent combinations tested, the original Tool-MAD setup (RAG + SEARCH) achieved the highest average accuracy, demonstrating its robustness and effectiveness.

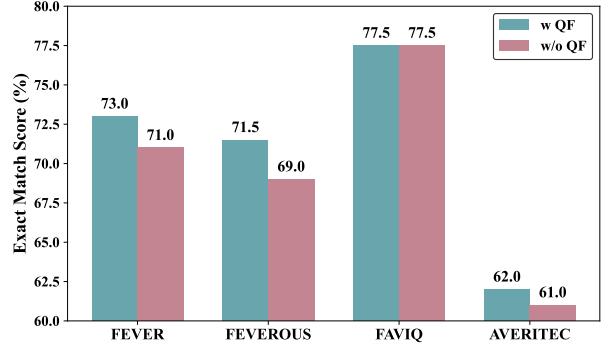


Fig. 9. Exact Match performance comparison with and without query formulation (QF) across four benchmark datasets.

In particular, configurations like RAG+RAG or SEARCH+SEARCH outperformed the VANILLA setup, but still fell short of the hybrid configuration (RAG+SEARCH). This suggests that relying on a single type of tool can lead to redundant retrievals and limit the diversity of evidence, thereby constraining overall performance improvements.

To further interpret these results, we note that the hybrid configuration benefits from the complementary characteristics of the two external tools. RAG-based retrieval provides semantically aligned, context-rich documents, while search-based retrieval offers broader coverage and access to rapidly updated information. These heterogeneous evidence sources allow the agents to explore distinct perspectives during the debate, reducing shared blind spots and producing more reliable conclusions. In contrast, homogeneous configurations tend to retrieve overlapping content, which limits the diversity of arguments and constrains potential performance gains.

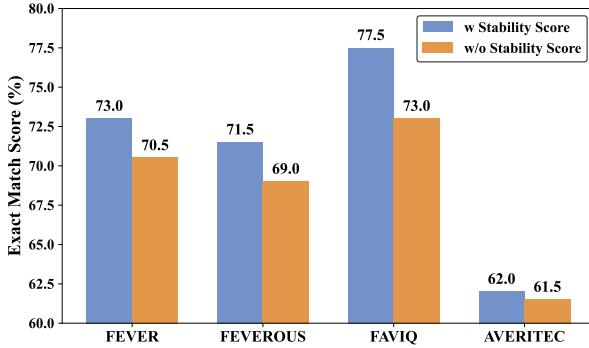


Fig. 10. Faithfulness and Answer Relevance scores across four datasets, based on the final predictions used for Exact Match evaluation, comparing models with and without scoring-based feedback.

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 8. 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.

This pattern suggests that although additional rounds help refine arguments and retrieve more relevant evidence, excessive debate depth may introduce unnecessary speculation or amplify intermediate reasoning errors. Such behavior aligns with observations in prior multi-step reasoning research, where prolonged interactions can lead to unstable or redundant argumentation. Therefore, three rounds provide sufficient opportunity for evidence refinement without causing degradation in the quality of the debate.

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 9, dynamic query formulation improves performance on all datasets except FAVIQ, where the score remains unchanged. The most notable improvement is observed on FEVER (+2.0) and FEVEROUS (+2.5), followed by a modest gain on AVERITEC (+1.0). These results demonstrate that formulating queries in response to the opponent’s argument helps uncover more relevant evidence and contributes to improved fact verification accuracy.

One possible explanation for this behavior is the difference in dataset characteristics. FEVER and FEVEROUS primarily contain entity-centric factual claims where emphasizing specific entities or relations during query reformulation leads to more targeted retrieval. In contrast, FAVIQ includes broader

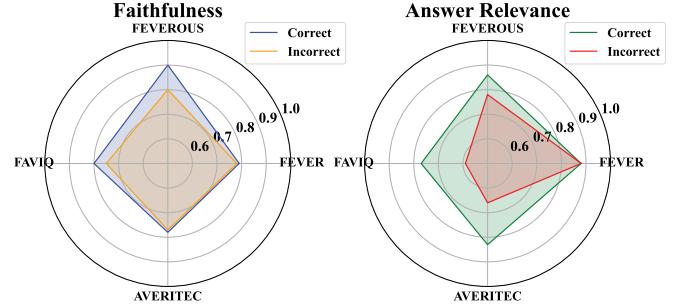


Fig. 11. The figure presents the average Faithfulness and Answer Relevance scores for both correct and incorrect answers across four datasets.

information-seeking questions that often require diverse or multi-hop evidence, meaning that the initial query already captures most of the relevant search space. As a result, further refinement has limited impact on FAVIQ, whereas datasets with more structured claim formulations benefit more from adaptive querying.

Effect of Stability Score We conducted ablation experiments to assess the effectiveness of incorporating faithfulness and answer relevance (stability score) as dynamic evaluation signals. In Tool-MAD, each agent’s response is evaluated at every debate round using these two metrics, and if either score falls below predefined thresholds, the round is considered inconclusive and is repeated. This mechanism enables the framework to dynamically filter out low-quality responses throughout the debate.

To isolate the impact of this score-based filtering, we compare the standard Tool-MAD configuration that utilizes faithfulness and answer relevance scores to guide debate progression (w/ Scoring Feedback) with a variant that skips score-based validation and progresses through all rounds regardless of response quality (w/o Scoring Feedback).

The results in Figure 10 highlight that scoring feedback consistently improves EM performance across all datasets, with the most significant gain observed on FAVIQ (+4.5%). These findings suggest that faithfulness and answer relevance function not merely as evaluation metrics but as crucial control signals that help maintain factual consistency and question relevance throughout multi-round debates.

A notable aspect of this trend is that FAVIQ benefits particularly strongly from stability-based filtering. Because many FAVIQ questions are open-ended and sensitive to topical alignment, answers that deviate slightly from the intended question are often incorrect even when they contain factually valid information. In such cases, answer relevance becomes a highly predictive indicator of correctness, while faithfulness prevents unsupported reasoning from propagating across rounds. Together, these signals help stabilize the debate process and reduce both topic drift and hallucinated content.

F. Analysis of Stability Score

Figure 11 compares the average Faithfulness and Answer Relevance scores across four datasets, based on whether the

agent's prediction was correct or incorrect. In all datasets, responses classified as correct consistently exhibit higher scores on both metrics, indicating a strong correlation between these scores and the Exact Match (EM) accuracy. These findings suggest that the scores can be used not only as post-hoc evaluation metrics but also as internal signals for determining answer correctness or filtering responses in real-time systems.

In particular, the FAVIQ and AVERITEC datasets show a pronounced gap in Answer Relevance scores between correct and incorrect responses, suggesting that alignment with the original user query plays a crucial role in predicting correctness for complex question answering tasks. These results highlight that Faithfulness and Answer Relevance serve as strong and reliable indicators for evaluating response quality and factual consistency. They can also contribute to quality control and final decision-making mechanisms in multi-agent debate systems.

V. DISCUSSION

A. Structural Effects of Query Rewriting

One of the most salient patterns observed in our analysis is that the gradual query rewriting performed throughout the debate plays a crucial role in improving verification accuracy. Initial queries typically take a broad, holistic form that mirrors the surface structure of the claim, but as the debate progresses, the agents refine their queries by decomposing attributes, restructuring entity relations, or explicitly specifying missing details. For instance, in the Trinidad and Tobago Guardian ownership case, the first-round query asked about the newspaper's overall "ownership history," which retrieved only partial information. In later rounds, the agents rewrote the query to isolate the key relation: "the relationship between Guardian Media Limited and ANSA McAL," which surfaced documents describing the hierarchical ownership chain. This shift transformed an initially ambiguous or contradictory prediction into a stable SUPPORTS decision. A similar phenomenon occurred in the Mel Ott RBI case. Early queries retrieved only season-level statistics, leading to NOT ENOUGH INFO (NEI), but rewriting the query into "total career RBIs" enabled retrieval of the exact aggregate value (1860), allowing the system to recover from partial evidence and converge on the correct label. These examples demonstrate that the multi-round query rewriting process is not merely repeated retrieval, but a structured mechanism that incrementally decomposes and reconstructs the claim, guiding the system toward increasingly precise evidence.

B. Structural Effects of Stability scores

A second notable characteristic identified in the analysis is the stabilizing effect of round-wise Stability scores, which evaluate each answer's faithfulness (document–answer consistency) and answer relevance (question–answer consistency). It is common for the RAG-based answer and the Search-based answer to propose different labels for the same claim. Rather than selecting based on majority or repetition, the system computes stability scores for both answers at every round and tends to choose the label whose supporting answer

maintains consistently higher semantic alignment with the retrieved evidence and the original question. For example, in the Stomatochaeta classification case, the RAG answer repeatedly predicted SUPPORTS while the Search answer predicted REFUTES. Across multiple rounds, however, the Search answer exhibited higher faithfulness scores and more stable question–answer relevance, prompting the model to adopt REFUTES as the final label. The Obispo first ascent case further illustrates this effect. Although the first round produced NEI due to insufficient evidence, subsequent rounds produced SUPPORTS answers with steadily increasing stability scores, eventually overriding the initial uncertainty and yielding a confident final decision. Because stability score directly measures whether an answer is grounded in retrieved evidence, it remains robust even under noisy retrieval conditions, assigning greater weight to evidence-aligned answers that consistently satisfy semantic constraints. Thus, stability score provides a meta-evaluative layer that re-assesses each candidate answer based on evidence quality and coherence, substantially enhancing convergence reliability compared to single-agent RAG systems.

C. Multi-Agent Debate as a Structural Exploration Mechanism

Finally, the analysis shows that the debate process exhibits an emergent capacity to explore structural properties of evidence, temporal shifts, and conflicting information even without an explicit moderator. A representative example is the Fort Myers Police Department case, where the Search tool retrieved up-to-date information indicating that the police chief had recently passed away, implying REFUTES, while the ground-truth label, anchored to an earlier timestamp, remained SUPPORTS. In such scenarios, the agents did not merely default to the labeled answer but instead engaged in a comparative evaluation of evidence recency, reliability, and contextual relevance. In other cases, the agents grounded their arguments in different retrieval sources and iteratively critiqued one another's evidence, adjusting their interpretations in later rounds. This behavior suggests that the system is not simply producing parallel answers but is performing a form of quasi-critical reasoning that assesses the coherence, conflict, and incompleteness of available evidence. Such patterns indicate that multi-agent debate has the potential to evolve beyond static fact verification toward a richer inference system capable of handling temporal drift, evidence conflicts, and multi-faceted claims, conditions that more closely resemble real-world reasoning tasks.

VI. CONCLUSION

In this work, we introduced Tool-MAD, a multi-agent debate framework that enhances factual verification by assigning distinct external tools to each agent and enabling adaptive, context-aware retrieval throughout the debate. Across diverse fact-verification and medical QA benchmarks, Tool-MAD achieved up to 35% performance gains over existing debate systems, with ablation studies confirming the importance of tool diversity, dynamic query formulation, and our Stability

Score, which functions as an internal control signal rather than a retrospective metric. These results demonstrate that combining structured debate with adaptive tool use significantly improves factual grounding and robustness. Tool-MAD's stable performance across domains and retrieval configurations indicates strong potential for extension to richer tool ecosystems and more advanced judge models. Overall, Tool-MAD provides a unified and evidence-sensitive framework for transparent and reliable multi-agent reasoning in high-stakes factual verification tasks.

REFERENCES

- [1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, “Language models are few-shot learners,” *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
- [2] R. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du *et al.*, “Lamda: Language models for dialog applications,” *arXiv preprint arXiv:2201.08239*, 2022.
- [3] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, A. Wahid, J. Tompson, Q. Vuong, T. Yu, W. Huang *et al.*, “Palm-e: An embodied multimodal language model,” 2023.
- [4] J. Li, J. Chen, R. Ren, X. Cheng, W. X. Zhao, J.-Y. Nie, and J.-R. Wen, “The dawn after the dark: An empirical study on factuality hallucination in large language models,” *arXiv preprint arXiv:2401.03205*, 2024.
- [5] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, “Chain-of-thought prompting elicits reasoning in large language models,” *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.
- [6] N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, and S. Yao, “Reflexion: Language agents with verbal reinforcement learning,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 8634–8652, 2023.
- [7] T. Liang, Z. He, W. Jiao, X. Wang, Y. Wang, R. Wang, Y. Yang, S. Shi, and Z. Tu, “Encouraging divergent thinking in large language models through multi-agent debate,” in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 17 889–17 904. [Online]. Available: <https://aclanthology.org/2024.emnlp-main.992/>
- [8] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch, “Improving factuality and reasoning in language models through multiagent debate,” in *Forty-first International Conference on Machine Learning*, 2023.
- [9] H. Wang, X. Du, W. Yu, Q. Chen, K. Zhu, Z. Chu, L. Yan, and Y. Guan, “Learning to break: Knowledge-enhanced reasoning in multi-agent debate system,” *Neurocomputing*, vol. 618, p. 129063, 2025.
- [10] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, “Retrieval-augmented generation for knowledge-intensive nlp tasks,” *Advances in neural information processing systems*, vol. 33, pp. 9459–9474, 2020.
- [11] S. Es, J. James, L. E. Anke, and S. Schockaert, “Ragas: Automated evaluation of retrieval augmented generation,” in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 2024, pp. 150–158.
- [12] M. Minsky, *Society of mind*. Simon and Schuster, 1986.
- [13] N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, and S. Yao, “Reflexion: language agents with verbal reinforcement learning,” in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 8634–8652. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/1b44b878bb782e6954cd888628510e90-Paper-Conference.pdf
- [14] J. C.-Y. Chen, S. Saha, and M. Bansal, “Reconcile: Round-table conference improves reasoning via consensus among diverse llms,” *arXiv preprint arXiv:2309.13007*, 2023.
- [15] Y. Chen, Q. Fu, Y. Yuan, Z. Wen, G. Fan, D. Liu, D. Zhang, Z. Li, and Y. Xiao, “Hallucination detection: Robustly discerning reliable answers in large language models,” in *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 2023, pp. 245–255.
- [16] Y. Zhao, J. Zhang, I. Chern, S. Gao, P. Liu, J. He *et al.*, “Felm: Benchmarking factuality evaluation of large language models,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 44 502–44 523, 2023.
- [17] S. Min, K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P. W. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi, “Factscore: Fine-grained atomic evaluation of factual precision in long form text generation,” *arXiv preprint arXiv:2305.14251*, 2023.
- [18] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou, “Self-consistency improves chain of thought reasoning in language models,” *arXiv preprint arXiv:2203.11171*, 2022.
- [19] P. Manakul, A. Liusie, and M. Gales, “Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models,” in *Proceedings of the 2023 conference on empirical methods in natural language processing*, 2023, pp. 9004–9017.
- [20] T. Schick, J. Dwivedi-Yu, R. Dessi, R. Raileanu, M. Lomeli, E. Hambro, L. Zettlemoyer, N. Cancedda, and T. Scialom, “Toolformer: Language models can teach themselves to use tools,” in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 68 539–68 551. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/d842425e4bf79ba039352da0f658a906-Paper-Conference.pdf
- [21] Y. Shen, K. Song, X. Tan, D. Li, W. Lu, and Y. Zhuang, “Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 38 154–38 180, 2023.
- [22] Y. Lu, H. Yu, and D. Khashabi, “Gear: Augmenting language models with generalizable and efficient tool resolution,” *arXiv preprint arXiv:2307.08775*, 2023.
- [23] A. M. Bran, S. Cox, O. Schilter, C. Baldassari, A. D. White, and P. Schwaller, “Chemcrow: Augmenting large-language models with chemistry tools,” *arXiv preprint arXiv:2304.05376*, 2023.
- [24] S. Gao, A. Fang, Y. Huang, V. Giunchiglia, A. Noori, J. R. Schwarz, Y. Ektefaie, J. Kondic, and M. Zitnik, “Empowering biomedical discovery with ai agents,” *Cell*, vol. 187, no. 22, pp. 6125–6151, 2024.
- [25] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, “Retrieval augmented language model pre-training,” in *International conference on machine learning*. PMLR, 2020, pp. 3929–3938.
- [26] J. Wang, X. Yi, R. Guo, H. Jin, P. Xu, S. Li, X. Wang, X. Guo, C. Li, X. Xu *et al.*, “Milvus: A purpose-built vector data management system,” in *Proceedings of the 2021 International Conference on Management of Data*, 2021, pp. 2614–2627.
- [27] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, “Fever: a large-scale dataset for fact extraction and verification,” *arXiv preprint arXiv:1803.05355*, 2018.
- [28] R. Aly, Z. Guo, M. Schlichtkrull, J. Thorne, A. Vlachos, C. Christodoulopoulos, O. Cocarascu, and A. Mittal, “Feverous: Fact extraction and verification over unstructured and structured information,” *arXiv preprint arXiv:2106.05707*, 2021.
- [29] J. Park, S. Min, J. Kang, L. Zettlemoyer, and H. Hajishirzi, “Faviq: Fact verification from information-seeking questions,” *arXiv preprint arXiv:2107.02153*, 2021.
- [30] M. Schlichtkrull, Z. Guo, and A. Vlachos, “Averitec: A dataset for real-world claim verification with evidence from the web,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 65 128–65 167, 2023.
- [31] D. Jin, E. Pan, N. Oufattolle, W.-H. Weng, H. Fang, and P. Szolovits, “What disease does this patient have? a large-scale open domain question answering dataset from medical exams,” *Applied Sciences*, vol. 11, no. 14, p. 6421, 2021.
- [32] Q. Jin, B. Dhingra, Z. Liu, W. W. Cohen, and X. Lu, “Pubmedqa: A dataset for biomedical research question answering,” *arXiv preprint arXiv:1909.06146*, 2019.
- [33] OpenAI, “Introducing gpt-4o: our fastest and most affordable flagship model,” <https://platform.openai.com/docs/guides/vision>, 2024, accessed: 2025-05-19.
- [34] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan *et al.*, “The llama 3 herd of models,” *arXiv preprint arXiv:2407.21783*, 2024.
- [35] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi *et al.*, “Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning,” *arXiv preprint arXiv:2501.12948*, 2025.
- [36] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, “Large language models are zero-shot reasoners,” *Advances in neural information processing systems*, vol. 35, pp. 22 199–22 213, 2022.
- [37] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, “React: Synergizing reasoning and acting in language models,” in *International Conference on Learning Representations (ICLR)*, 2023.

- [38] J. Priem, H. Piwowar, and R. Orr, “Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts,” *arXiv preprint arXiv:2205.01833*, 2022.
- [39] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “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)*, 2019, pp. 4171–4186.