Exploring Health Misinformation Detection with Multi-Agent Debate

Chih-Han Chen*
National Taiwan University
chen.eric0208@gmail.com

Chen-Han Tsai
HTC DeepQ
maxwell_tsai@htc.com

Yu-Shao Peng
HTC DeepQ
ys_peng@htc.com

Abstract

Fact-checking health-related claims has become increasingly critical as misinformation proliferates online. Effective verification requires both the retrieval of high-quality evidence and rigorous reasoning processes. In this paper, we propose a two-stage framework for health misinformation detection: Agreement Score Prediction followed by Multi-Agent Debate. In the first stage, we employ large language models (LLMs) to independently evaluate retrieved articles and compute an aggregated agreement score that reflects the overall evidence stance. When this score indicates insufficient consensus—falling below a predefined threshold—the system proceeds to a second stage. Multiple agents engage in structured debate to synthesize conflicting evidence and generate well-reasoned verdicts with explicit justifications. Experimental results demonstrate that our two-stage approach achieves superior performance compared to baseline methods, highlighting the value of combining automated scoring with collaborative reasoning for complex verification tasks.

1 Introduction & Related Work

The proliferation of health-related content on digital platforms poses significant challenges to ensuring accurate medical information reaches the public. Verifying health claims is critical for safeguarding public well-being, as false or misleading information can cause substantial harm to individual and population health. Despite the vast volume of health content available online, only a small fraction is supported by robust scientific evidence, underscoring the urgent need for automated verification systems.

In open-domain fact-checking, traditional methods predominantly rely on BERT-based architectures (Devlin et al., 2019). Pipeline-based systems employ BERT models to retrieve relevant evidence sentences, followed by a classification module to predict claim veracity. Joint systems perform evidence retrieval and veracity prediction simultaneously within a unified model. While conceptually straightforward, these approaches require predefined knowledge databases and necessitate training encoder-based models from scratch (Vladika et al., 2024), limiting their flexibility and scalability.

The emergence of large language models (LLMs) has introduced new paradigms. Tian et al. (2024) deploy web retrieval agents to gather evidence dynamically, enabling LLMs to assess sufficiency and render verdicts. Singal et al. (2024) integrate retrieval-augmented generation (RAG) with in-context learning (ICL) for veracity prediction. Vladika et al. (2025) propose multi-turn LLM interactions that iteratively generate questions, retrieve evidence, and reason about claim validity. However, these approaches typically lack explicit evidence filtering mechanisms, relying directly on outputs from web search tools or dense retrieval models.

Recent work has explored multi-agent frameworks for fact-checking. Hong et al. (2025) leverage multiple agents to evaluate evidence quality and determine veracity, with provisions for regathering evidence when necessary. Hu et al. (2025), Liang et al. (2024), and (Liu et al., 2025) adopt Multi-Agent Debate (MAD) frameworks to enhance reasoning robustness and mitigate degenerate reasoning patterns.

Building upon these advances, we propose a two-stage multi-agent debate framework for health misinformation detection. Our approach first employs LLMs to retrieve and evaluate high-quality articles, computing an aggregated agreement score. When evidence exhibits significant disagreement—indicated by a score below a predefined threshold—the system initiates a structured multi-agent debate. Through iterative argumenta-

*Work done while interning at HTC DeepQ.

Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications, pages 16–21
December 23, 2025 ©2025 Association for Computational Linguistics

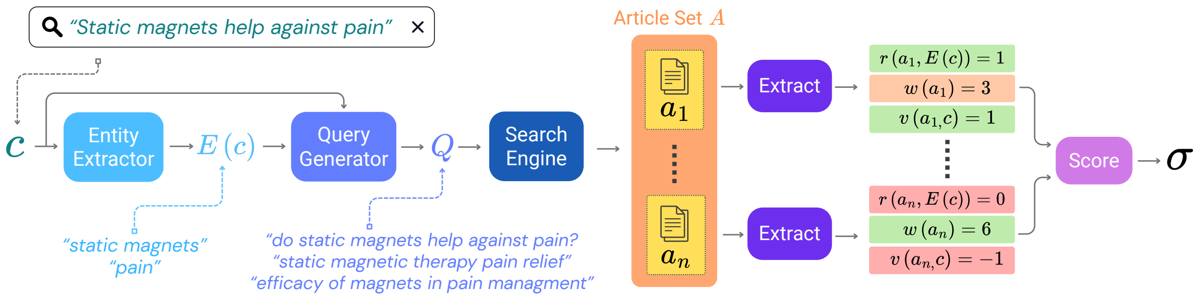

Figure 1: **Agreement Score Prediction (Stage 1).** For a given claim c, entities $E(c)$ are extracted and passed alongside c to a query generator to generate search queries Q. Articles relating to Q are collected into an article set A. We extract topic relevance r, article weights w, and article verdict v for each article $a \in A$. The results are aggregated, resulting in the final agreement score σ.

tion, agents collaboratively analyze conflicting evidence to produce well-justified verdicts grounded in explicit reasoning.

2 Methodology

In this section, we detail the implementation of our proposed two-stage health misinformation detection algorithm. The first stage takes a claim as input and retrieves a set of articles relating to the claim. Each article is classified as to whether it *Supports* or *Refutes* the claim, and the predictions are aggregated. When the agreement among predictions is high, the veracity of the claim is determined by majority vote. In the case of low agreement, we initiate the second stage multi-turn debate. Two opposing agents are provided with supporting evidence collected during the first stage, and a judge agent supervises the debate process until the claim's veracity can be determined. The details of each stage are presented in the following.

2.1 Agreement Score Prediction

Figure 1 illustrates the first stage framework of our approach. For a given claim c, we first extract a set of entities $E(c)$ from c using an LLM. The entities are keywords or phrases from c that the claim is focused on. The claim c and entities $E(c)$ are then provided to an LLM to generate a set of queries Q. Each query $q \in Q$ is sent to a search engine for article retrieval. The article sets retrieved from each query are de-duplicated and merged to form the article set A.

Given the obtained queries Q, entities $E(c)$, and article set A, we prompt an LLM to extract the following information from each article $a \in A$. Specifically, we look for:

1. **Topic Relevance:** Check whether the article a contains content relevant for all entities in $E(c)$. We define this relevance as $r(a, E(c)) \in \{0, 1\}$, where $r(a, E(c)) = 1$ if the article contains content relevant for all entities in $E(c)$ and $r(a, E(c)) = 0$ otherwise.

2. **Attribute Assessment:** Evaluate whether article a contains the following attributes: *Problem Statement*, *Experimental Setup*, *Findings*, *Statistical Significance*, *Limitations*, and *Results*. These 6 attributes reflect the structure of modern scientific publications. Specifically, an article that coverts the 6 attributes are often more thorough in its claims. We define the article weight as:

$$w(a) = \sum_{\alpha \in \text{Attributes}} \mathbf{1}[\alpha \in a] \in \{0, 1, \ldots, 6\}$$

where $\mathbf{1}[\cdot]$ is the indicator function for whether attribute α is in article a.

3. **Article Verdict:** Determine whether the contents of the article a *support* or *refute* the claim c. We denote $v(a, c) \in \{-1, 1\}$ where $v(a, c) = 1$ indicates *support* and $v(a, c) = -1$ indicates *refute*.

We then compute the *agreement score* $\sigma(c, A) \in [-1, 1]$ for claim c and article set A as:

$$\sigma(c, A) = \frac{1}{Z} \sum_{a \in A} r(a, E(c)) \cdot w(a) \cdot v(a, c),$$

where

$$Z = \sum_{a \in A} r(a, E(c)) \cdot w(a)$$

is the normalizing constant. We consider the case where $Z \neq 0$ by assuming quality relevant

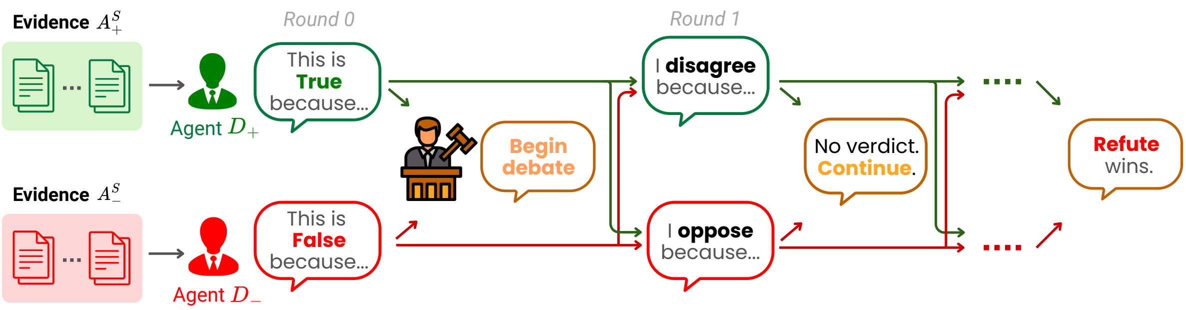

Figure 2: **Multi-Agent Debate (Stage 2).** Articles from the first Agreement Score Prediction stage are organized into supporting and refuting evidence sets A_+^S and A_-^S, which are provided to agents D_+ and D_-, respectively. Each agent begins with an opening statement based on their evidence set, after which the judge initiates the debate. In each round, agents review their opponent's argument before providing a counterargument. After each round, the judge determines whether sufficient information exists to reach a verdict. If not, the debate continues for another round. The process concludes when the judge reaches a final verdict.

articles to be available within the search engine results.

We introduce a threshold $\tau > 0$ to quantify the *level of agreement* among the retrieved articles. If $|\sigma| \geq \tau$, this indicates that most articles consistently *support* or *refute* the claim. When such high level of agreement exists, the first stage directly outputs *support* for $\sigma \geq \tau$ and *refute* for $\sigma \leq -\tau$.

Conversely, an agreement score $|\sigma| < \tau$ indicates a significant level of disagreement among the articles. In this case, we pass the results to the second stage for debate.

2.2 Multi-Agent Debate

Figure 2 illustrates the second stage framework of our approach. We employ a multi-agent debate framework based on the work by Liang et al. (2024). The debate framework involves three agents: the **Support Agent** D_+, **Refute Agent** D_-, and **Judge Agent** J. Evidence is first prepared using the results from the first stage before initiating the debate.

Evidence Preparation: Given the article set A, we select two disjoint subsets A_+ and A_- from A such that:

$$A_+ = \{a \in A \mid v(a,c) = +1,\ r(a, E(c)) = 1\},$$
$$A_- = \{a \in A \mid v(a,c) = -1,\ r(a, E(c)) = 1\}.$$

Articles in A_+ and A_- are ranked in descending order using $w(a)$, and we limit each set to contain an equal number of articles. For each article in the remaining sets, we prompt an LLM using the claim c to extract passages from the original text that *supports* or *refutes* claim c along with its reason. We concatenate the LLM responses from all articles in sets A_+ and A_- into A_+^s and A_-^s. We denote A_+^s and A_-^s as the *supporting* and *refuting evidence* throughout the debate process.

Opening Statement: The support agent D_+ and refute agent D_- begins with an opening statement by presenting the evidence in A_+^S and A_-^S. We denote the outputs of the support and refute agents as

$$S_+^{(0)} = D_+(A_+^s), \quad S_-^{(0)} = D_-(A_-^s).$$

Each agent also maintains a conversation history H. Following the opening statement, we initialize each agent's history as

$$H_+^{(0)} = \{S_+^{(0)}\}, \quad H_-^{(0)} = \{S_-^{(0)}\}.$$

The judge agent's history is initialized using the opening statements given by the two debate agents

$$H_J^{(0)} = \{S_+^{(0)}, S_-^{(0)}\}.$$

Next, the judge initiates the debate process, and we proceed to the first round of debate.

Debate Process: In every debate round, each agent responds to the opposing agent's statement $S^{(i-1)}$ using its past conversation history $H^{(i-1)}$. The outputs of the support and refute agent from the i-th round are given as

$$S_+^{(i)} = D_+\big(S_-^{(i-1)}, H_+^{(i-1)}\big),$$
$$S_-^{(i)} = D_-\big(S_+^{(i-1)}, H_-^{(i-1)}\big).$$

The debate agent's histories are updated by concatenating the opposing agent's response along with the current response

$$H_+^{(i)} = H_+^{(i-1)} \oplus S_-^{(i-1)} \oplus S_+^{(i)},$$
$$H_-^{(i)} = H_-^{(i-1)} \oplus S_+^{(i-1)} \oplus S_-^{(i)}.$$

The judge agent J takes the response from both agents along with its own history $H_{\text{J}}^{(i-1)}$, and decides whether sufficient information exists to reach a verdict. Specifically,

$$\theta^{(i)} = J\left(S_+^{(i)},\ S_-^{(i)},\ H_{\text{J}}^{(i-1)}\right)$$

where $\theta^{(i)} \in \{\texttt{support}, \texttt{refute}, \texttt{continue}\}$. If the judge agent believes an argument is compelling enough, the verdict $\theta^{(i)} \in \{\texttt{support}, \texttt{refute}\}$ is returned. If neither argument is sufficiently convincing, the judge agent outputs $\theta^{(i)} = \texttt{continue}$, and the debate continues for another round.

The judge's history is also updated by appending the debate agent responses

$$H_{\text{J}}^{(i)} = H_{\text{J}}^{(i-1)} \oplus S_+^{(i)} \oplus S_-^{(i)}.$$

To prevent indefinitely long debates, we limit the process to a maximum of M rounds, after which the judge must reach a verdict $\theta^{(M)} \in \{\texttt{support}, \texttt{refute}\}$ based on the debate history.

3 Experiments and Setup

3.1 Datasets

We consider the following health-related datasets for our experiments.

SciFact (Wadden et al., 2020) contains expert-written biomedical claims derived from medical paper abstracts. We use the development subset, consisting of 188 claims: 124 supported and 64 refuted.

TREC-Health (Pugachev et al., 2023) is constructed from the TREC 2019 Decision Track (Abualsaud et al., 2020) and the TREC 2021 Health Misinformation Track (Clarke et al., 2021), both of which target challenges in search engine results related to health misinformation. The dataset includes 113 consumer health questions, of which 61 are supported and 52 are refuted.

HealthFC (Vladika et al., 2024) consists of everyday health-related claims spanning diverse topics. We use a subset of 327 claims: 202 supported and 125 refuted.

3.2 Metrics

We report macro-precision, macro-recall, and macro-F1 as evaluation metrics. These are standard in fact-checking tasks, as they provide a balanced analysis of prediction performance across labels.

3.3 Baseline Algorithms

We consider WebAgent (Tian et al., 2024) and StepByStep (Vladika et al., 2025) as benchmark algorithms. Among them, StepByStep represents the current state-of-the-art in health-related fact-checking. For fairness, all methods, including ours, use the Brave search engine (Brave Software, Inc.) and GPT-4o (OpenAI, 2024) as the underlying LLM. Each algorithm is executed three times, and we report the best performance.

For our framework, we set the parameters as follows: entity set size $|E(c)| = 2$, query set size $|Q| = 5$, article set size $|A| = 10$, agreement threshold $\tau = 0.7$, and debate round limit $M = 5$.

3.4 Comparison Results & Analysis

The experimental results are shown in Table 1. Our first-stage-only method achieves better performance comparable to WebAgent, although StepByStep remains challenging to surpass.

When the second-stage debate mechanism is incorporated, our approach yields substantial improvements over the first-stage-only variant: F1 scores increase by $+3.1$ on TREC-Health and $+8.1$ on HealthFC. This demonstrates that, in cases of uncertain agreement among retrieved articles, the debate mechanism enables more effective reasoning and leads to stronger overall performance.

Compared to StepByStep, our two-stage pipeline achieves higher F1 performance by $+0.8$ on TREC-Health and $+1.4$ on HealthFC. Notably, our method maintains a balance between precision and recall, whereas StepByStep tends to favor high recall at the expense of precision.

Table 2 reports results on the high-agreement subset. High coverage and strong performance in this setting show that the first stage can reliably resolve many claims. However, when evidence is sparse or contradictory, the second-stage debate provides the additional reasoning needed, underscoring its critical role in the framework.

4 Conclusion

We proposed a two-stage framework for health misinformation detection that combines agreement score prediction with multi-agent debate. The first stage leverages weighted agreement scoring to resolve claims directly, while the second stage provides explainable reasoning through debate.

Experiments on three health datasets demonstrate consistent improvements over strong base-

Method	SciFact			TREC-Health			HealthFC		
	P	R	F1	P	R	F1	P	R	F1
WebAgent (Tian et al., 2024)	80.1	83.2	80.6	76.2	75.6	75.7	78.0	78.3	78.1
StepByStep (Vladika et al., 2025)	**86.1**	**89.5**	**87.8**	69.9	**95.1**	80.6	72.6	**91.6**	81.0
Ours (1st stage only)	84.9	86.1	85.5	**83.8**	78.2	78.3	76.9	73.4	74.3
Ours (1st stage + 2nd stage)	82.4	85.3	83.1	81.3	81.5	**81.4**	**82.1**	82.7	**82.4**

Table 1: Performance comparison across three datasets (SciFact, TREC-Health, and HealthFC) using macro precision (P), recall (R), and F1 score. Best results are in **bold**.

	SciFact	TREC-Health	HealthFC
Coverage	64.9%	50.1%	58.1%
F1 Score	92.0	88.6	84.0

Table 2: Results on the high-agreement subset. *Coverage (%)* denotes the proportion of claims settled without debate in the first stage, while *F1 Score* reports the score for those claims.

lines, including gains of +0.8 F1 on TREC-Health and +1.4 F1 on HealthFC, with a better balance between precision and recall. These results underscore the value of integrating evidence consistency with structured debate, advancing reliable and explainable health misinformation detection.

Limitations

While our two-stage framework achieves strong performance, it also entails certain limitations. First, as the approach relies on LLMs, the debate judge may still be affected by model biases or occasional hallucinations. Second, the multi-agent design requires multiple API calls, introducing extra computational cost; however, this cost is modest compared to the performance gains. Finally, our current evaluation is limited to binary-labeled datasets. Extending the framework to more nuanced settings, such as incorporating a *Not Enough Information* class, represents a promising direction for future work.

References

Mustafa Abualsaud, Christina Lioma, Maria Maistro, Mark D. Smucker, and Guido Zuccon. 2020. Overview of the trec 2019 decision track.

Brave Software, Inc. Brave search api documentation. `https://api.search.brave.com/res/v1/web/search`. Accessed: 2025-08-21.

Charles L. A. Clarke, Saira Rizvi, Mark D. Smucker, Maria Maistro, and G. Zuccon. 2021. Overview of the trec 2021 health misinformation track. In *Text Retrieval Conference*.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Spencer Hong, Meng Luo, and Xinyi Wan. 2025. Emulate: A multi-agent framework for determining the veracity of atomic claims by emulating human actions. *Preprint*, arXiv:2505.16576.

Wentao Hu, Wengyu Zhang, Yiyang Jiang, Chen Jason Zhang, Xiaoyong Wei, and Li Qing. 2025. Removal of hallucination on hallucination: Debate-augmented RAG. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15839–15853, Vienna, Austria. Association for Computational Linguistics.

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.

Yuhan Liu, Yuxuan Liu, Xiaoqing Zhang, Xiuying Chen, and Rui Yan. 2025. The truth becomes clearer through debate! multi-agent systems with large language models unmask fake news. *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

OpenAI. 2024. Gpt-4o. `https://openai.com/index/hello-gpt-4o`. Accessed: 2025-08-21.

Alexander Pugachev, Ekaterina Artemova, Alexander Bondarenko, and Pavel Braslavski. 2023. Consumer health question answering using off-the-shelf components. In *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part II*, page 571–579, Berlin, Heidelberg. Springer-Verlag.

Ronit Singal, Pransh Patwa, Parth Patwa, Aman Chadha, and Amitava Das. 2024. Evidence-backed fact checking using RAG and few-shot in-context learning with

LLMs. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 91–98, Miami, Florida, USA. Association for Computational Linguistics.

Jacob-Junqi Tian, Hao Yu, Yury Orlovskiy, Tyler Vergho, Mauricio Rivera, Mayank Goel, Zachary Yang, Jean-François Godbout, Reihaneh Rabbany, and Kellin Pelrine. 2024. Web retrieval agents for evidence-based misinformation detection. In *First Conference on Language Modeling*.

Juraj Vladika, Ivana Hacajova, and Florian Matthes. 2025. Step-by-step fact verification system for medical claims with explainable reasoning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 805–816, Albuquerque, New Mexico. Association for Computational Linguistics.

Juraj Vladika, Phillip Schneider, and Florian Matthes. 2024. HealthFC: Verifying health claims with evidence-based medical fact-checking. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8095–8107, Torino, Italia. ELRA and ICCL.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.