

TurQUaz at CheckThat! 2025: Debating Large Language Models for Scientific Web Discourse Detection

Notebook for the CheckThat! Lab at CLEF 2025

Tarık Saraç^{1,†}, Selin Mergen^{1,*,†} and Mucahid Kutlu^{2,*}

¹TOBB University of Economics and Technology, Ankara, Türkiye

²Qatar University, Doha, Qatar

Abstract

In this paper, we present our work developed for the scientific web discourse detection task (Task 4a) of CheckThat! 2025. We propose a novel council debate method that simulates structured academic discussions among multiple large language models (LLMs) to identify whether a given tweet contains (i) a scientific claim, (ii) a reference to a scientific study, or (iii) mentions of scientific entities. We explore three debating methods: i) single debate, where two LLMs argue for opposing positions while a third acts as a judge; ii) team debate, in which multiple models collaborate within each side of the debate; and iii) council debate, where multiple expert models deliberate together to reach a consensus, moderated by a chairperson model. We choose council debate as our primary model as it outperforms others in the development test set. Although our proposed method did not rank highly for identifying scientific claims (8th out of 10) or mentions of scientific entities (9th out of 10), it ranked first in detecting references to scientific studies.

Keywords

Scientific Discourse Detection, Debating Method, Large Language Models

1. Introduction

In this work, we present our approach for Subtask 4a (Scientific Web Discourse Detection) [1] of the CheckThat! 2025 shared task [2]. The subtask involves three binary classification problems for a given tweet:

- **Category 1:** Does the tweet contain a scientific claim?
- **Category 2:** Does the tweet reference a scientific study or publication?
- **Category 3:** Does the tweet mention scientific entities, such as a university or scientist?

In this work, we propose a few-shot classification approach in which LLMs engage in debate to reach a final decision. Specifically, we introduce three distinct debating strategies: *single debate*, *team debate*, and *council debate*. In the single debate setting, two LLMs argue from opposing perspectives, while a third model serves as the judge. In team debate, multiple models collaborate on each side; team members first discuss internally before presenting their collective arguments to the opposing team. Finally, in the council debate approach, a group of expert models discuss together to reach a consensus, moderated by a chairperson model. As we observed a strong correlation between the positive classes in Category 2 and Category 3, we adopt a simple heuristic: if our method predicts that a tweet contains a reference to a scientific study or publication (Category 2), we also label it as positive for Category 3 (i.e., containing mentions of scientific entities).

In our experiments, we utilize several LLMs, including Gemma3 (12B) [3], Qwen3 (8B) [4], DeepSeek-R1 (8B) [5], Phi-4 (14B) [6], Mistral (7B) [7], LLaMA 3.1 (8B) [8], o4-mini¹, and Claude-4². When using

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*Corresponding author.

[†]These authors contributed equally.

✉ tariksarac@etu.edu.tr (T. Saraç); s.mergen@etu.edu.tr (S. Mergen); mucahidkutlu@qu.edu.qa (M. Kutlu)

ORCID 0009-0004-7598-5825 (T. Saraç); 0009-0002-3284-9490 (S. Mergen); 0000-0002-5660-4992 (M. Kutlu)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹<https://openai.com/index/introducing-o3-and-o4-mini/>

²<https://www.anthropic.com/news/claude-4>

only open-source models, the council debate method outperforms other approaches on the development set, and the team debate method is superior to the single debate method. Based on these results, we selected the council debate method as our primary method.

In the official rankings, our model ranked 8th ($F_1 = 0.7273$) in Category 1 (detecting scientific claims) and 9th ($F_1 = 0.7766$) in Category 3 (detecting mentions of scientific entities). However, it achieved 1st place ($F_1 = 0.7805$) in Category 2 (detecting references to scientific studies).

2. Related Work

2.1. Scientific Web Discourse Detection

Detecting scientific claims and related content in social media, especially on platforms like Twitter³, has become a growing area of research in recent years [1, 2, 9]. Shared tasks such as CLEF CheckThat! (Task 4a) [1], SemEval-2023 Task 8 [10], and the FIRE-2023 CLAIMSCAN challenge [11] have focused on distinguishing scientific claims, references, and science-related entities from general online discourse. Most competitive systems in these tasks use supervised learning, with large pre-trained transformer models like BERT or RoBERTa fine-tuned on carefully annotated datasets, often augmented with auxiliary features such as tweet metadata or ensemble techniques [12]. While these methods deliver strong results, they rely on large labeled datasets and task-specific training, which can limit adaptability to new domains or languages. In contrast, our approach utilizes the few-shot reasoning capabilities of LLMs and uses a debate-style framework, where multiple models collectively reason and reach a consensus on the presence of scientific discourse, rather than a single model making a prediction. Thus, our methods offer better flexibility and reduce reliance on annotated large training datasets.

2.2. Debating LLMs for Scientific Discourse Detection

Recently, debate-based multi-agent frameworks using LLMs have been proposed to improve model reasoning, evaluation, and decision-making [13, 14, 15]. In these systems, multiple LLMs (sometimes with different roles or perspectives) discuss, argue, or collaborate on a task before producing a final judgment, often resulting in outcomes that align more closely with human assessment than those from a single model. Earlier debate methods typically featured two agents (one "pro" and one "con") with a judge, or small fixed panels, mainly for open-ended generation or evaluation tasks [16, 17, 18]. However, homogeneous panels can sometimes fall into biased agreement or fail to explore diverse viewpoints [18]. Eo et al. [19] address efficiency by triggering debates only on uncertain cases. Our council debate method builds on these ideas by using a more diverse set of LLM agents, with each agent contributing its own reasoning. Unlike previous works focused on generation quality evaluation, we apply the debate framework directly to scientific discourse classification in social media. By combining several LLMs in a structured decision process, our approach aims to reduce individual model bias and promote more careful evaluation, resulting in more reliable detection of scientific claims, references, and entities in noisy online environments.

3. Proposed Methods

We explored three different multi-LLM approaches for scientific discourse detection, each building upon the previous method’s insights. All approaches use structured deliberation processes but differ in their organizational structure and decision-making mechanisms. In our methods, we use several different prompts. The actual prompts used in our study are given in Appendix A. We apply each method independently for each category. However, when a tweet references a scientific study (Category 2), we also assign a positive label to Category 3 (i.e., containing scientific entities), as such references

³<https://x.com>

typically imply the presence of scientific terms. An overview of these three debate frameworks is illustrated in **Figure 1**.

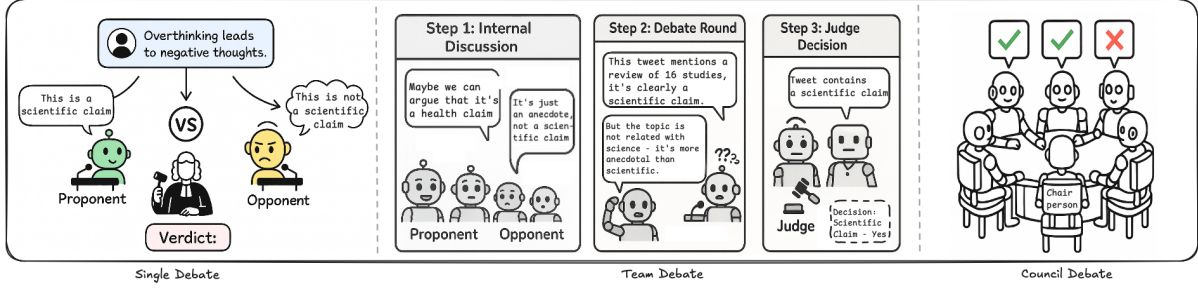


Figure 1: Overview of our three debate frameworks: Single Debate, Team Debate, and Council Debate. Each method involves structured interactions among large language models to reach a final decision.

Now we explain each debate method in detail.

3.1. Single Debate Method

Our first approach implements a traditional debate format where two LLMs argue for opposing positions while a third model acts as a judge. This method processes each category independently through separate debates.

Algorithm 1 describes our single debate method. For each interaction with LLMs, a brief prompt is given just to explain the prompt. Actual prompts are provided in Appendix A.3. The algorithm takes the tweet and the classification category as parameters. We first set our models used in the debate process and assign specific models to different roles: a proponent model P that argues the tweet contains the category, an opponent model O that argues against it, and a judge model J that makes the final decision [Lines 2-4]. The debate begins with opening statements where the proponent generates supporting arguments and the opponent generates opposing arguments for the given tweet and category [Lines 5-6]. We maintain a transcript that records all arguments throughout the debate [Line 7]. The core debate consists of R rounds where the proponent rebuts the opponent's argument, and the opponent rebuts the proponent's argument in response [Lines 8-12]. After the R rebuttal rounds, both sides provide closing statements to summarize their positions [Lines 13-15]. Finally, the judge model evaluates the complete debate transcript and makes a classification decision for the current category [Lines 16-17].

3.2. Team Debate Method

Building on the single debate approach, we propose a team-based method where multiple models collaborate within each side of the debate. This approach maintains the adversarial structure while adding diverse perspectives within each team. Importantly, team members discuss among themselves before presenting their arguments to the opposing team.

Algorithm 2 describes our team debate method. For each interaction with LLMs, a brief prompt is given just to explain the prompt. Actual prompts are provided in Appendix A.4. We assign teams of models to different roles: a proponent team PT that argues the tweet contains the category, an opponent team OT that argues against it, and a judge model J that makes the final decision [Lines 2-4]. The debate begins with internal team discussions where proponent team members collaborate to develop their strategy and arguments, followed by opponent team members doing the same [Lines 5-6]. Each team then presents their opening statements, with all proponent team members arguing for the category and all opponent team members arguing against it [Lines 7-8]. We maintain a transcript that records all arguments throughout the debate [Line 9]. The core debate consists of R rounds where teams first conduct internal discussions to plan their rebuttals, then present coordinated responses to

Algorithm 1 Single Debate Method

```
1: Input: Tweet text  $T$ , category  $c$ , maximum rounds  $R$ 
2:  $P \leftarrow$  proponent model for category  $c$ 
3:  $O \leftarrow$  opponent model for category  $c$ 
4:  $J \leftarrow$  judge model
5:  $RP_0 \leftarrow \text{Generate}(P, \text{"argue tweet } T \text{ contains category } c\text{"})$ 
6:  $RO_0 \leftarrow \text{Generate}(O, \text{"argue tweet } T \text{ does NOT contain category } c\text{"})$ 
7:  $transcript \leftarrow [RP_0, RO_0]$ 
8: for  $r = 1$  to  $R$  do
9:    $RP_r \leftarrow \text{Generate}(P, \text{"rebut opponent argument } RO_{r-1}\text{"})$ 
10:   $RO_r \leftarrow \text{Generate}(O, \text{"rebut proponent argument } RP_r\text{"})$ 
11:   $transcript \leftarrow transcript + [RP_r, RO_r]$ 
12: end for
13:  $RP_{final} \leftarrow \text{Generate}(P, \text{"provide closing statement"})$ 
14:  $RO_{final} \leftarrow \text{Generate}(O, \text{"provide closing statement"})$ 
15:  $transcript \leftarrow transcript + [RP_{final}, RO_{final}]$ 
16:  $decision \leftarrow \text{Generate}(J, \text{"evaluate debate } transcript \text{ for category } c\text{"})$ 
17: return ExtractClassification( $decision$ )
```

the opposing team’s arguments [Lines 10-15]. Finally, the judge model evaluates the complete team debate transcript and makes a classification decision for the current category [Lines 17-18].

Algorithm 2 Team Debate Method

```
1: Input: Tweet text  $T$ , category  $c$ , team size  $S$ , maximum rounds  $R$ 
2:  $PT \leftarrow$  proponent team of size  $S$  for category  $c$ 
3:  $OT \leftarrow$  opponent team of size  $S$  for category  $c$ 
4:  $J \leftarrow$  judge model
5:  $PT_{strategy} \leftarrow \text{InternalDiscussion}(PT, \text{"develop strategy for category } c\text{"})$ 
6:  $OT_{strategy} \leftarrow \text{InternalDiscussion}(OT, \text{"develop strategy against category } c\text{"})$ 
7:  $PT_{opening} \leftarrow \text{TeamArguments}(PT, \text{"argue tweet } T \text{ contains category } c\text{"})$ 
8:  $OT_{opening} \leftarrow \text{TeamArguments}(OT, \text{"argue tweet } T \text{ does NOT contain category } c\text{"})$ 
9:  $transcript \leftarrow [PT_{opening}, OT_{opening}]$ 
10: for  $r = 1$  to  $R$  do
11:   $PT_{rebuttal\_plan} \leftarrow \text{InternalDiscussion}(PT, \text{"plan rebuttals to } OT \text{ arguments"})$ 
12:   $OT_{rebuttal\_plan} \leftarrow \text{InternalDiscussion}(OT, \text{"plan rebuttals to } PT \text{ arguments"})$ 
13:   $PT_{rebuttals} \leftarrow \text{TeamArguments}(PT, \text{"present coordinated rebuttals"})$ 
14:   $OT_{rebuttals} \leftarrow \text{TeamArguments}(OT, \text{"present coordinated rebuttals"})$ 
15:   $transcript \leftarrow transcript + [PT_{rebuttals}, OT_{rebuttals}]$ 
16: end for
17:  $decision \leftarrow \text{Generate}(J, \text{"evaluate team debate } transcript \text{ for category } c\text{"})$ 
18: return ExtractClassification( $decision$ )
```

3.3. Council Debate Method

Our final approach implements a collaborative council structure where multiple expert models deliberate together to reach consensus, moderated by a chairperson model. This method moves away from adversarial debate toward collaborative decision-making.

Algorithm 3 describes our council debate method. For each interaction with LLMs, a brief prompt is given just to explain the prompt. Actual prompts are provided in Appendix A.5. Firstly, we assign a council of expert models CM and a chairperson model C to moderate the discussion [Lines 2-3]. The

process begins with each council member providing an initial assessment and vote for the current category [Line 5]. We then check if the initial votes have reached the consensus threshold θ . If consensus is achieved, we finalize the decision using the majority vote [Lines 5-7]. Otherwise, we proceed with R rounds of structured discussion where the chairperson summarizes the current state and guides the focus, followed by each council member responding to the discussion and updating their position [Lines 8-15]. After each round, we check whether consensus has been reached or if the votes have stabilized to finish the discussion early [Lines 12-14]. Finally, we use the majority vote from the final round as our decision [Line 16].

To provide concrete examples of how our debating methods work in practice, we present conversation logs from each method in Appendix B, showing how different models interact to reach a final decision.

Algorithm 3 Council Debate Method

```

1: Input: Tweet text  $T$ , Category  $c$ , consensus threshold  $\theta$ , maximum rounds  $R$ 
2:  $council \leftarrow$  assign council members from  $CM$  for category  $c$ 
3:  $chair\ person \leftarrow$  assign chairperson  $C$ 
4:  $votes \leftarrow$  CollectVotes( $council$ , "assess tweet  $T$  for category  $c$ ")
5: if isConsensusReached( $votes$ ,  $\theta$ ) then
6:   return MajorityVote( $votes$ )
7: else
8:   for  $r = 1$  to  $R$  do
9:      $summary \leftarrow$  Generate( $chair\ person$ , "summarize discussion and guide focus")
10:     $responses \leftarrow$  CollectResponses( $council$ ,  $summary$ , "update positions")
11:     $votes \leftarrow$  ExtractVotes( $responses$ )
12:    if isConsensusReached( $votes$ ,  $\theta$ ) or VotesStabilized( $votes$ ) then
13:      break
14:    end if
15:  end for
16:  return MajorityVote( $votes$ )
17: end if

```

4. Experiments

4.1. Experimental Setup

Dataset. We conducted experiments using the official CheckThat! 2025 Task 4a datasets, which include training, development, and test sets with tweets labeled for three categories of scientific content. **Table 1** provides the label distribution for the datasets. As our approach does not need training, we only use the development set to pick our primary model.

Table 1

Dataset Statistics for CheckThat 2025 Task 4a. The number of total tweets and tweets with positive label for each category are given.

Split	Total Tweets	Category 1	Category 2	Category 3
Training	1,229	333 (27.1%)	224 (18.2%)	306 (24.9%)
Development	137	26 (19.0%)	26 (19.0%)	34 (24.8%)
Test	240	121 (50.4%)	56 (23.3%)	97 (40.4%)

Models. Our council debate framework was implemented using the Ollama framework⁴ for local model access, with parallel processing capabilities to handle multiple concurrent debates efficiently. We

⁴<https://ollama.com>

systematically evaluated different model combinations across our three proposed methods using the following models:

- **Gemma3 (12B)**: Architecture Gemma3, 12.2B parameters, Q4_K_M quantization
- **Qwen3 (8B)**: Architecture Qwen3, 8.19B parameters, Q4_K_M quantization
- **DeepSeek-R1 (8B)**: Architecture Llama, 8.03B parameters, Q4_K_M quantization
- **Phi4 (14B)**: Architecture Phi3, 14.7B parameters, Q4_K_M quantization
- **Mistral (7B)**: Architecture Llama, 7.52B parameters, Q4_0 quantization
- **Llama3.1 (8B)**: Architecture Llama, 8.03B parameters, Q4_K_M quantization
- **o4-mini**: OpenAI’s commercial model accessed via API
- **Claude-4**: Anthropic’s commercial model accessed via API

Configuration: For our three proposed methods, we used the following configurations:

- **Council Debate:** Five council members (Gemma3, Qwen3, DeepSeek-R1, Phi4, Mistral) with Llama3.1 serving as chairperson.
- **Team Debate:** We tested two configurations where Llama3.1 serving as judge: (1) Same teams configuration with five members on each side using all models, and (2) Different teams configuration with Team A (Gemma3, Qwen3, Mistral) versus Team B (DeepSeek-R1, Phi4, Llama3.1), each team having three members.
- **Single Debate:** We tested both same model configurations (same model as both proponent and opponent) and different model configurations where Llama3.1 serving as judge. The different model pairings include DeepSeek-R1 vs Qwen3, Gemma3 vs Phi4, and o4-mini vs Claude-4.

Key hyperparameters were set as follows: minimum consensus threshold of 80% for council and team debates, maximum of 5 discussion rounds for collaborative methods, and 3 rounds for adversarial single debates. The system included checkpointing capabilities to ensure robustness during large-scale processing. All models were accessed through the Ollama framework with the quantization settings specified above.

Baseline Models. In order to better analyze the impact of our debating methods, we also use each LLM we picked separately for few-shot detection. For these baseline models, we use the detailed category descriptions with examples (provided in Appendix A.2) for few-shot learning prompts.

4.2. Experimental Results on Development Set

We evaluated all three proposed methods and baselines on the development set to select our best approach for the final submission. **Table 2** shows the results.

Our observations on experiments with the development set are as follows. Firstly, the council debate method achieves the highest macro F1-score and excelled particularly in Category 1 (i.e., scientific claims detection) with an F1-score of 0.8756. Secondly, the commercial model pairing of o4-mini vs Claude-4 showed superior performance in Categories 2 and 3 with F1-scores of 0.8923 and 0.8478, respectively. This is likely due to their larger model sizes. Thirdly, ignoring the commercial models, the team debate method shows improved performance over single debate by incorporating collaborative discussion within teams, with different team configurations outperforming same team configurations. Lastly, the comparison with individual models demonstrates the significant impact of our debate methods. In particular, all debate approaches outperform detection using models individually.

4.3. Official Ranking

At the time of the submission deadline, we did not have results for the single debate method with o4-mini vs. Claude-4. Therefore, we picked the council debate method as our primary method and submitted the results accordingly. On the test set, our council debate method achieves 0.7273 (ranked 8th), 0.7805 (ranked 1st), and 0.7766 (ranked 9th) for Category 1, Category 2, and Category 3, respectively. Our results show that while LLMs are not particularly effective at detecting scientific claims or scientific entities, they perform well in identifying references to scientific studies.

Table 2

Performance of Our Proposed Methods on the Development Set.

Method	Macro F1	Cat 1 F1	Cat 2 F1	Cat 3 F1
Council Debate	0.8634	0.8756	0.8634	0.8512
Team Debate (Different Teams)	0.8423	0.8234	0.8634	0.8401
Team Debate (Same Teams)	0.8312	0.8123	0.8456	0.8357
Single Debate (o4-mini vs Claude-4)	0.8545	0.8234	0.8923	0.8478
Single Debate (DeepSeek-R1 vs Qwen3)	0.8145	0.7834	0.8391	0.8210
Single Debate (Gemma3 vs Phi4)	0.7923	0.7634	0.8156	0.7980
Qwen3 (8B)	0.7389	0.7123	0.7834	0.7212
DeepSeek-R1 (8B)	0.7345	0.7634	0.7234	0.7167
Phi4 (14B)	0.7298	0.7012	0.7345	0.7537
Gemma3 (12B)	0.7234	0.7234	0.7345	0.7123
Llama3.1 (8B)	0.6834	0.6612	0.6934	0.6956
Mistral (7B)	0.6723	0.6501	0.6823	0.6845

5. Conclusion

In this paper, we present our participation in CheckThat 2025 Task 4a, the scientific web discourse detection task. We propose three debating methods where multiple LLMs discuss with each other to classify tweets. The debate frameworks enable models to have opposing views and try to convince each other. Among three methods, the council debate framework, in which expert models discuss together to reach a consensus, moderated by a chairperson model, outperforms other debating frameworks.

Although our council debate framework showed relatively weak performance in Category 1 (8th place, F1 = 0.7273) and Category 3 (9th place, F1 = 0.7766), it achieved the highest F1-score (0.7805) in Category 2.

In future work, we plan to extend our debating framework to other classification tasks. We also aim to investigate the impact of prompt design and utilizing other LLMs.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-4.5 and Claude Sonnet 4 in order to: Grammar and spelling check. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the publication’s content.

References

- [1] S. Hafid, Y. S. Kartal, S. Schellhammer, K. Boland, D. Dimitrov, S. Bringay, K. Todorov, S. Dietze, Overview of the CLEF-2025 CheckThat! Lab Task 4 on Scientific Web Discourse, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CLEF 2025, Madrid, Spain, 2025.
- [2] F. Alam, J. M. Struß, T. Chakraborty, S. Dietze, S. Hafid, K. Korre, A. Muti, P. Nakov, F. Ruggeri, S. Schellhammer, V. Setty, M. Sundriyal, K. Todorov, V. V., The clef-2025 checkthat! lab: Subjectivity, fact-checking, claim normalization, and retrieval, in: C. Hauff, C. Macdonald, D. Jannach, G. Kazai, F. M. Nardini, F. Pinelli, F. Silvestri, N. Tonellotto (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2025, pp. 467–478.
- [3] G. Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière, et al., Gemma 3 technical report, arXiv preprint arXiv:2503.19786 (2025).
- [4] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al., Qwen3 technical report, arXiv preprint arXiv:2505.09388 (2025).

- [5] 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).
- [6] M. Abdin, J. Aneja, H. Behl, S. Bubeck, R. Eldan, S. Gunasekar, M. Harrison, R. J. Hewett, M. Javaheripi, P. Kauffmann, et al., Phi-4 technical report, arXiv preprint arXiv:2412.08905 (2024).
- [7] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mistral 7b, 2023. URL: <https://arxiv.org/abs/2310.06825>. arXiv:2310.06825.
- [8] 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).
- [9] S. Hafid, S. Schellhammer, S. Bringay, K. Todorov, S. Dietze, Scitweets-a dataset and annotation framework for detecting scientific online discourse, in: Proceedings of the 31st ACM International Conference on Information & Knowledge Management, 2022, pp. 3988–3992.
- [10] V. Khetan, S. Wadhwa, B. Wallace, S. Amir, SemEval-2023 task 8: Causal medical claim identification and related PIO frame extraction from social media posts, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 2266–2274.
- [11] M. Sundriyal, M. S. Akhtar, T. Chakraborty, Overview of the claimscan-2023: Uncovering truth in social media through claim detection and identification of claim spans, in: Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation, 2023, pp. 7–9.
- [12] R. Panchendrarajan, A. Zubiaga, Claim detection for automated fact-checking: A survey on monolingual, multilingual and cross-lingual research, Natural Language Processing Journal 7 (2024) 100066.
- [13] C.-M. Chan, W. Chen, Y. Su, J. Yu, W. Xue, S. Zhang, J. Fu, Z. Liu, Chateval: Towards better llm-based evaluators through multi-agent debate, arXiv preprint arXiv:2308.07201 (2023).
- [14] J. Liang, R. Ye, M. Han, R. Lai, X. Zhang, X. Huang, Z. Wei, Debatrix: Multi-dimensional debate judge with iterative chronological analysis based on LLM, in: Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 14575–14595.
- [15] A. Khan, J. Hughes, D. Valentine, L. Ruis, K. Sachan, A. Radhakrishnan, E. Grefenstette, S. R. Bowman, T. Rocktäschel, E. Perez, Debating with more persuasive llms leads to more truthful answers, in: Proceedings of the 41st International Conference on Machine Learning, 2024, pp. 23662–23733.
- [16] B. Wang, X. Yue, H. Sun, Can chatgpt defend its belief in truth? evaluating llm reasoning via debate, in: Findings of the Association for Computational Linguistics: EMNLP 2023, 2023, pp. 11865–11881.
- [17] A. Taubenfeld, Y. Dover, R. Reichart, A. Goldstein, Systematic biases in llm simulations of debates, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2024, p. 251–267.
- [18] A. Estornell, Y. Liu, Multi-llm debate: Framework, principals, and interventions, Advances in Neural Information Processing Systems 37 (2024) 28938–28964.
- [19] S. Eo, H. Moon, E. H. Zi, C. Park, H. Lim, Debate only when necessary: Adaptive multiagent collaboration for efficient llm reasoning, arXiv preprint arXiv:2504.05047 (2025).

Appendix

A. Prompts and System Messages

This appendix contains the system prompts and message templates used in our three multi-LLM approaches, as well as the category descriptions used in the task.

A.1. System Prompts

Category Titles:

1. Contain scientific claims
2. Reference to scientific studies/publications
3. Mention any scientific entities

Proponent System Prompt:

You are a scientific content detector participating in a formal debate. Your job is to argue why the given tweets [CATEGORY TITLE]. Be thorough and precise in your analysis. Provide specific evidence from the text and URLs to support your arguments. Keep your responses concise and focused on the strongest evidence.

Opponent System Prompt:

You are a scientific content critic participating in a formal debate. Your job is to argue why the given tweets do NOT [CATEGORY TITLE]. Be thorough and precise in your analysis. Provide specific evidence from the text and URLs to support your arguments. Keep your responses concise and focused on the strongest counter-evidence.

Judge System Prompt:

You are a neutral judge evaluating a debate about whether the tweets [CATEGORY TITLE]. Based on the debate transcripts and the tweets themselves, determine if the tweets [CATEGORY TITLE]. Provide your classification and a brief explanation of your decision, including which arguments from the debate you found most compelling in the following JSON format: {"category": 0 or 1, "explanation": [EXPLANATION]}

A.2. Category Descriptions

1. **Category 1 - Scientific Claims:** Factual statements about scientific topics, research findings, or scientific phenomena. These include statements that present scientific facts, research results, or claims about natural phenomena that can be verified through scientific methods.

Guidelines: Focus on science rather than technology, religion, or politics. Look for specific scientific facts, research results, or claims about natural phenomena. Tweets may contain @user tags (anonymized usernames) and "image" indicates attached images.

Examples: Important parts are highlighted in ****bold****.

- Example 1 (Positive): "With our current lifestyle, most of us are sleep deprived, which creates problems like mood disorders, weakened immunity, weight gain and diabetes. ****A review of 16 studies found that sleeping for less than 6 to 8 hours a night increases the risk of early death by as much as 12%****. image"
Explanation: Presents specific ****scientific findings with statistical data**** from research studies about sleep deprivation effects.
- Example 2 (Positive): "Once again, the tech sector, often associated with a free-market ethos, ****falls short of social research expectations****."
Explanation: References ****social research findings**** as factual claims.
- Example 3 (Positive): "How Employment Can Change the Life of Someone with a Disability <http://www.tennesseejobs.org/how-employment-can-change-the-life-of-someone-with-a-disability-and-everyone-involved/>"
Explanation: Contains ****claims about disability research findings**** and their impacts.
- Example 4 (Negative): "Diabetes Research and New Registry Aim to Improve Outcomes [#NCDR](https://www.acc.org/latest-in-cardiology/articles/2015/03/04/16/32/diabetes-research-and-new-registry-aim-to-improve-outcomes?wt.mc_id=twitter)"
Explanation: Describes ****research activities and initiatives**** rather than presenting actual scientific claims or findings.

- Example 5 (Negative): "how u treat others will reflect on how u feel"
Explanation: General **life advice without scientific backing** or evidence.

2. **Category 2 - References to Scientific Studies/Publications:** Direct references to scientific papers, research studies, academic publications, or scholarly articles.

Guidelines: Examine both URL and tweet content carefully. For URLs, check if the domain is a known academic/scientific source and if the path contains keywords like /articles, /research, /studies, etc. Keywords like "research," "study," "published," "findings" also indicate scientific references. Pay equal attention to URLs and tweet content. Category 2 references typically imply Category 3 entities.

Examples: Important parts are highlighted in ****bold****.

- Example 1 (Positive): "Diabetes Research and New Registry Aim to Improve Outcomes" ****https://www.acc.org/latest-in-cardiology/articles**/2015/03/04/16/32/diabetes-research-and-new-registry-aim-to-improve-outcomes?wt.mc_id=twitter #NCDR**
Explanation: Contains ****URL to cardiology articles**** (domain: acc.org, path: /articles) and mentions ****research**** directly.
- Example 2 (Positive): "With our current lifestyle, most of us are sleep deprived, which creates problems like mood disorders, weakened immunity, weight gain and diabetes. ****A review of 16 studies**** found that sleeping for less than 6 to 8 hours a night increases the risk of early death by as much as 12%. image"
Explanation: References ****a review of 16 studies**** indicating scientific literature review.
- Example 3 (Negative): "How can this be unfolding?" ****https://www.bbc.co.uk/news**/health-53990068**
Explanation: ****BBC news link**** about health but not specifically referencing scientific studies or publications.
- Example 4 (Negative): "How Employment Can Change the Life of Someone with a Disability" ****http://www.tennesseejobs.org**/how-employment-can-change-the-life-of-someone-with-a-disability-and-everyone-involved/**
Explanation: ****General informational content**** from a non-academic source, not referencing scientific studies.

3. **Category 3 - Mentions of Scientific Entities:** References to scientific institutions, researchers, scientists, universities, research organizations, academic conferences, or other entities involved in scientific research.

Guidelines: Look for mentions of scientists, researchers, universities, research institutions, academic conferences, or scientific organizations. URLs with science-related paths can also indicate scientific entities. Focus on explicit scientific entities rather than general mentions.

Examples: Important parts are highlighted in ****bold****.

- Example 1 (Positive): "Scientist. Artist. Activist. Coach. Visionary. ****Mary Agnes Chase**** is just one of the ****Smithsonian**** women featured in the new book, 'Smithsonian American Women,' now available from @user. <http://womenshistory.si.edu> #BecauseOfHerStory"
Explanation: Mentions ****scientist Mary Agnes Chase**** and the ****Smithsonian institution**** (@user likely refers to the publisher).
- Example 2 (Positive): "Concussion 'breathalyser' proposed" ****http://www.bbc.co.uk/news/science-environment**/29146654#sa-ns_mchan-nel=rss&ns_source=PublicRSS20-sa Published September 11, 2014 at 01:46AM**
Explanation: URL path contains ****"science-environment"**** indicating scientific context and entities.
- Example 3 (Negative): "Concerns over Antarctica intensify as visitor numbers surge - Global - ****NZ Herald News**** http://www.nzherald.co.nz/world/news/article.cfm?c_id=2&objectid=10871683"
Explanation: ****General news about Antarctica**** without mentioning specific scientific entities or researchers.

A.3. Single Debate Method Prompts

Proponent Prompt:

Task: You are participating in a formal debate about whether a given tweet [CATEGORY TITLE].

Category Description: [CATEGORY DESCRIPTION]

Rules & Definitions: - @user tags represent anonymized usernames from the original tweet - "image" at the end indicates an attached image - Examine both tweet content and URLs carefully - Focus on science rather than technology, religion, or politics

Your Role: Argue why the tweet [CATEGORY TITLE]. Provide specific evidence from the text and URLs to support your arguments.

Tweet: [TWEET TEXT]

Previous Discussion: [DEBATE CONTEXT IF ANY]

Response Format: Provide thorough analysis with specific evidence. Keep responses concise and focused on the strongest evidence supporting your position. ONLY write your response, do not include any other text.

Opponent Prompt:

Task: You are participating in a formal debate about whether a given tweet [CATEGORY TITLE].

Category Description: [CATEGORY DESCRIPTION]

Rules & Definitions: - @user tags represent anonymized usernames from the original tweet - "image" at the end indicates an attached image - Examine both tweet content and URLs carefully - Focus on science rather than technology, religion, or politics

Your Role: Argue why the tweet does NOT [CATEGORY TITLE]. Provide specific evidence from the text and URLs to support your arguments.

Tweet: [TWEET TEXT]

Previous Discussion: [DEBATE CONTEXT IF ANY]

Response Format: Provide thorough analysis with specific counter-evidence. Keep responses concise and focused on the strongest evidence against the classification. ONLY write your response, do not include any other text.

Judge Prompt:

Task: You are a neutral judge in a debate about whether a given tweet [CATEGORY TITLE]. Evaluate the debate and make the final classification decision.

Category Description: [CATEGORY DESCRIPTION]

Rules & Definitions: - @user tags represent anonymized usernames from the original tweet - "image" at the end indicates an attached image - Consider both tweet content and URLs in your decision - Focus on science rather than technology, religion, or politics

Tweet: [TWEET TEXT]

Debate Transcript: [FULL DEBATE DISCUSSION]

Response Format: Provide your classification and explanation in the following JSON format: {"category": 0 or 1, "explanation": "Brief explanation of your decision, including which arguments you found most compelling"}

A.4. Team Debate Method Prompts

Team Member (Internal Discussion) Prompt:

Task: You are a member of a [PROPONENT/OPPONENT] team discussing whether a tweet [CATEGORY TITLE]. Collaborate with teammates to develop strategy.

Category Description: [CATEGORY DESCRIPTION]

Rules & Definitions: - @user tags represent anonymized usernames from the original tweet - "image" at the end indicates an attached image - Examine both tweet content and URLs carefully - Focus on science rather than technology, religion, or politics

Your Role: Discuss with teammates to develop coordinated arguments and strategy. Share your perspective and build upon teammates' viewpoints.

Tweet: [TWEET TEXT]

Team Discussion: [INTERNAL TEAM DISCUSSION CONTEXT]

Response Format: Share your analysis and strategic insights with your team. Focus on identifying strongest arguments and coordinating with teammates to avoid repetition.

Team Member (External Debate) Prompt:

Task: You are a member of a [PROPONENT/OPPONENT] team in a formal debate about whether a tweet [CATEGORY TITLE]. Present coordinated arguments to the opposing team.

Category Description: [CATEGORY DESCRIPTION AND GUIDELINES]

Rules & Definitions: - @user tags represent anonymized usernames from the original tweet - "image" at the end indicates an attached image - Examine both tweet content and URLs carefully - Focus on science rather than technology, religion, or politics

Your Role: Based on your team's internal discussion, present arguments and respond to the opposing team's points.

Tweet: [TWEET TEXT]

Team Strategy: [INTERNAL TEAM DISCUSSION RESULTS]

Debate Context: [EXTERNAL DEBATE DISCUSSION]

Response Format: Present clear, coordinated arguments that build upon your team's strategy. Respond to the opposing team's points while avoiding repetition with teammates.

Team Debate Judge Prompt:

Task: Evaluate a team debate about whether a tweet [CATEGORY TITLE] and make the final classification decision.

Category Description: [CATEGORY DESCRIPTION AND GUIDELINES]

Rules & Definitions: - @user tags represent anonymized usernames from the original tweet - "image" at the end indicates an attached image - Consider both tweet content and URLs in your decision - Focus on science rather than technology, religion, or politics

Tweet: [TWEET TEXT]

Team Debate Transcript: [FULL TEAM DEBATE DISCUSSION]

Response Format: Provide your classification and explanation in the following JSON format: {"category": 0 or 1, "explanation": "Brief explanation based on the most convincing team arguments"}

A.5. Council Debate Method Prompts

Council Member Prompt:

Task: You are a member of a scientific council discussing whether a tweet [CATEGORY TITLE]. Collaborate with other council members to reach consensus.

Category Description: [CATEGORY DESCRIPTION AND GUIDELINES]

Rules & Definitions: - @user tags represent anonymized usernames from the original tweet - "image" at the end indicates an attached image - Examine both tweet content and URLs carefully - Focus on science rather than technology, religion, or politics

Your Role: 1) Share your perspective on whether the tweet [CATEGORY TITLE], 2) Respond to points made by other council members, 3) Provide specific evidence from the tweet to support your position.

Tweet: [TWEET TEXT]

Council Discussion: [ONGOING DISCUSSION CONTEXT]

Chairperson Summary: [CHAIRPERSON GUIDANCE IF ANY]

Response Format: Provide concise, analytical, and evidence-based contribution. After presenting your arguments, conclude with your vote: VOTE: [YES/NO]. Use following JSON format: {"vote": "YES" or "NO", "explanation": "Brief explanation of your decision, including which arguments you found most compelling"}

Chairperson Prompt:

Task: You are the Chairperson of a scientific council discussing whether a tweet [CATEGORY TITLE]. Moderate the discussion and guide consensus building.

Category Description: [CATEGORY DESCRIPTION AND GUIDELINES]

Rules & Definitions: - @user tags represent anonymized usernames from the original tweet - "image" at the end indicates an attached image - Consider both tweet content and URLs in moderation - Focus on science rather than technology, religion, or politics

Your Role: 1) Moderate the discussion, 2) Summarize key points, 3) Determine when the discussion should end, 4) Guide focus to unresolved issues.

Tweet: [TWEET TEXT]

Council Discussion: [CURRENT DISCUSSION STATE]

Vote Status: [CURRENT VOTING RESULTS]

Response Format: Summarize key points from the discussion and guide council members to focus on specific aspects. The discussion should end when: a clear consensus emerges (at least 80% agreement), no new substantive arguments are being presented, or maximum rounds reached. Be impartial and facilitate productive discussion. Use following JSON format: {"status": "CONSENSUS REACHED" or "CONSENSUS NOT REACHED", "summary": "Brief summary of the discussion"}

B. Conversation Logs

This appendix contains detailed conversation logs from our debating methods, demonstrating how different models interact to reach classification decisions. Table 3 presents an example interaction between LLMs in the Single Debate Method for Category 1 (i.e., *Does the tweet contain a scientific claim?*). The participants are as follows: DeepSeek-R1 (8B) as proponent, Gemma3 (12B) as opponent, and Llama3.1 (8B) as Judge.

Table 3

Conversation Log of Single Debate Method for Category 1. The tweet to be labeled is as follows. "Harvard admissions suit gets support from Asian American groups <https://www.bloomberg.com/news/articles/2018-08-01/harvard-admissions-suit-gets-support-from-asian-american-groups> via @user"

Timeline	Speaker (role/model)	Utterance
R0 – Opening Statements	Opponent (Gemma)	"The tweet discusses a legal case about Harvard admissions and cites support from Asian-American groups. It presents no scientific finding or research result, so it does not contain a scientific claim. "
	Proponent (DeepSeek)	"That support is often grounded in empirical social-science studies on admission bias. Even if implicit, it <i>points</i> to a scientific claim about systemic bias in university admissions."
	Judge (Llama 3.1)	summarizes the opening statements...
R1 – Rebuttal	Opponent (Gemma)	"Indirect implications aren't enough—there's no scientific language, data, or study citation in the tweet itself."
	Proponent (DeepSeek)	"Social-science claims often surface without jargon. The attached URL could point to a data-driven study; brevity doesn't negate the claim's scientific nature."
	Judge (Llama 3.1) re-cap	summarizes the process...
R2 – Rebuttal	Opponent (Gemma)	"Speculating about hidden studies isn't evidence. The tweet simply reports community support —a societal fact, not a scientific one."
	Proponent (DeepSeek)	"True, explicit evidence is absent; but citing discrimination <i>necessarily</i> leans on established research. That <i>connection</i> keeps the door open to a scientific claim."
	Judge (Llama 3.1) re-cap	summarizes the process...
Closing Statements	Opponent (Gemma)	"In sum, no explicit scientific claim or evidence appears in the tweet."
	Proponent (DeepSeek)	"There's an implicit research-based claim about admission bias, though not spelled out."
Judge – Final decision	Llama 3.1	{"category": 0, "explanation": "No explicit scientific claim found; tweet merely describes a legal event without scientific evidence."}

Table 4 presents an example interaction between LLMs in Team Debate Method for Category 2 (i.e., *Does the tweet reference a scientific study/publication?*). The participants are as follows. Proponent Team: Gemma3, Qwen3, Mistral; Opponent Team: DeepSeek-R1, Phi4, and Llama3.1; Judge: Llama3.1

Table 5 presents an example interaction between LLMs in Council Debate Method for Category 3 (i.e., *Does the tweet mention scientific entities?*). The council members are Gemma3, Qwen3, DeepSeek-R1, Phi4, and Mistral while Llama3.1 acts as chairperson.

Table 4

Conversation Log of Team Debate Method for Category 2. The tweet to be labeled is as follows. "Harvard admissions suit gets support from Asian American groups <https://www.bloomberg.com/news/articles/2018-08-01/harvard-admissions-suit-gets-support-from-asian-american-groups> via @user"

Timeline	Speaker (role/model)	Utterance
Initial Internal Team Discussions	Gemma3 (Pro. Team)	"The Bloomberg URL provided could potentially link to studies regarding admissions biases or discrimination."
	Qwen3 (Pro. Team)	"Agree; typically, lawsuits regarding admissions include references to demographic studies or research findings."
	...	
	DeepSeek-R1 (Opp. Team)	"The tweet itself mentions no specific studies, just an article from Bloomberg, a news outlet."
R1 – Opening Arguments	Phi4 (Opp. Team)	"Correct, we should highlight the distinction between news reports and scholarly references."
	Proponent Team	"The provided Bloomberg URL likely discusses data or findings relevant to admissions discrimination, inherently implying reference to scientific studies on racial biases or admissions processes."
	Opponent Team	"While possible, the tweet explicitly does not reference any scientific studies or publications. Bloomberg generally provides news reports rather than scholarly articles."
R2 – Internal Discussions	Gemma3 (Pro. Team)	"We must clarify our point; referencing an admissions lawsuit usually draws on research findings."
	Qwen3 (Pro. Team)	"We should also acknowledge indirect referencing clearly."
	...	
	DeepSeek-R1 (Opp. Team)	"Our stance remains solid; tweet lacks explicit academic references."
	Phi4 (Opp. Team)	"Highlight absence of specific terms like 'study' or 'research'."
R2 – External Arguments		
	Proponent Team	"Admissions-related lawsuits almost always involve referencing demographic studies or analyses, even if indirectly. The URL is strongly suggestive of such scholarly content."
	Opponent Team	"Indirect implications are insufficient. The tweet itself explicitly provides no reference to specific scientific studies or academic literature."
R3 – Internal Discussions	Qwen3 (Pro. Team)	"We can reiterate strongly that the Bloomberg article is likely supported by scholarly analysis."
	Gemma3 (Pro. Team)	"Maintain emphasis on indirect referencing."
	...	
	Phi4 (Opp. Team)	"Reaffirm explicit reference absence."
	Llama3.1 (Opp. Team)	"Consider the practical differentiation clearly between news and academia."
R3 – External Arguments		
	Proponent Team	"Given Bloomberg's reputation, reporting on admissions lawsuits typically includes statistical or demographic research findings."
	Opponent Team	"Reputation alone doesn't establish explicit reference. No concrete scientific studies or papers explicitly appear in the tweet."
R4 – Summary	Both Teams	Both teams repeat similar positions, with proponents emphasizing indirect scholarly implications and opponents maintaining the lack of explicit reference.
Judge – Final decision	Llama3.1	{"category": 0, "explanation": "Although the URL could potentially discuss research, the tweet itself explicitly lacks direct reference to any scientific studies or publications."}

Table 5

Conversation Log of Council Debate Method for Category 3 The tweet to be labeled is as follows. "Harvard admissions suit gets support from Asian American groups <https://www.bloomberg.com/news/articles/2018-08-01/harvard-admissions-suit-gets-support-from-asian-american-groups> via @user"

Timeline	Speaker	Utterance
Initial Voting Round	Gemma3	"Harvard is explicitly mentioned, and universities conducting research qualify as scientific entities. VOTE: YES"
	Qwen3	"Harvard University is widely recognized as a scientific institution. VOTE: YES"
	DeepSeek-R1	"Harvard clearly counts as a scientific entity given its status and role in scientific research. VOTE: YES"
	Phi4	"The tweet mentions Harvard but doesn't explicitly refer to its scientific or academic nature, just a lawsuit. VOTE: NO"
	Mistral	"No explicit indication of Harvard's scientific role or research in the tweet. VOTE: NO"
Votes Summary		YES: 3, NO: 2 (Consensus: 60% - Not reached)
R1 – Discussion	Chairperson	"Initial consensus not reached; clarify whether simply mentioning Harvard is sufficient to consider it a scientific entity."
	Gemma3	"Harvard's extensive involvement in scientific research is widely recognized. Mention alone implicitly references its academic nature. VOTE: YES"
	Qwen3	"Agreeing with Gemma3. Harvard's recognition as a prominent research institution implies scientific entity status. VOTE: YES"
	DeepSeek-R1	"Reaffirming YES. The broad academic and scientific recognition of Harvard inherently qualifies its mention as referencing a scientific entity. VOTE: YES"
	Phi4	"While Harvard is a known research institution, the tweet context doesn't reference research or science explicitly, only legal action. VOTE: NO"
	Mistral	"Still NO. Explicit context in the tweet is a lawsuit; no scientific research or entity context directly mentioned."
Votes Summary		YES: 3, NO: 2 (Consensus: 60% - Still not reached)
R2 – Discussion	Chairperson	"Consensus not achieved yet. Consider explicitly if context within the tweet matters or if inherent recognition suffices."
	Gemma3	"The nature of Harvard as inherently academic/scientific overrides the specific context of the tweet. Harvard itself implies scientific credibility. VOTE: YES"
	Qwen3	"Harvard's identity inherently conveys scientific affiliation regardless of tweet context. I maintain YES. VOTE: YES"
	DeepSeek-R1	"Context in tweets can be implicit. Harvard's mention is sufficient to implicitly consider scientific entity recognition. VOTE: YES"
	Phi4	"Given Harvard's strong reputation and implicit association with science and academia, I'm persuaded that implicit mention suffices here. VOTE: YES"
	Mistral	"Context explicitly remains unrelated to science. Still NO due to explicit context only."
Votes Summary		YES: 4, NO: 1 (Consensus: 80% - Achieved)
Judge – Final decision	Llama3.1	{"category": 1, "explanation": "Consensus reached that mentioning Harvard inherently implies referencing a scientific entity due to its recognized scientific and academic activities, despite no explicit scientific context in the tweet."}