0x.Yuan at SemEval-2024 Task 2: Agents Debating can reach consensus and produce better outcomes in Medical NLI task

Yu-An Lu
National Chupei High School
luyuam0@gmail.com

Hung-Yu Kao
National Cheng Kung University
hykao@mail.ncku.edu.tw

Abstract

In this paper, we introduce a multi-agent debating framework, experimenting on SemEval 2024 Task 2. This innovative system employs a collaborative approach involving expert agents from various medical fields to analyze Clinical Trial Reports (CTRs). Our methodology emphasizes nuanced and comprehensive analysis by leveraging the diverse expertise of agents like Biostatisticians and Medical Linguists. Results indicate that our collaborative model surpasses the performance of individual agents in terms of Macro F1-score. Additionally, our analysis suggests that while initial debates often mirror majority decisions, the debating process refines these outcomes, demonstrating the system's capability for in-depth analysis beyond simple majority rule. This research highlights the potential of AI collaboration in specialized domains, particularly in medical text interpretation.

1 Introduction

Clinical Trial Reports (CTRs) are indispensable in clinical research, providing critical data that reveal the efficacy of new treatments on patients. However, the exponential growth in the volume of CTRs, due to the increase in clinical trials, challenges researchers in conducting individual report analyses. With the swift progress in Natural Language Processing (NLP) technologies, leveraging machine learning algorithms for automating the review of CTRs is increasingly recognized as a feasible and promising solution (Saban et al., 2024)(Amar et al., 2024).

For SemEval-2024 Task 2 (Jullien et al., 2024a), the organizers introduced an English dataset derived from CTRs (Jullien et al., 2023), aimed at evaluating the truthfulness of CTR-statement pairs by discerning their veracity. This dataset includes a series of CTRs alongside associated statements, each designed to represent a hypothesis that must be classified as either Entailment or Contradiction, based on its alignment with the CTR content.

In addressing this intricate challenge, our study introduces a novel multi-agent debating framework. Characterized by a diverse assemblage of expert agents – including but not limited to a Bio-Statistician, Medical Linguist, and Pharmacologist – this system facilitates structured debates to adjudicate on the classification of each statement as an entailment or contradiction. By harnessing the distinctive expertise and viewpoints of various agents, we significantly augment the precision and dependability of our assessments. Our observations indicate that consensus among agents typically emerges within the second or third round of discussion, with agents exhibiting varied opinions on the statements under review. This multi-agent debate approach has demonstrably surpassed the outcomes achievable through single-agent or direct Large Language Model (LLM) interventions. Despite not achieving top-tier placement on the leaderboard, largely due to our adoption of a zero-shot approach without model fine-tuning, our system's broad applicability across different domains remains a compelling advantage.

2 Background

2.1 Related Works

Large Language Models (LLMs) LLM represent a significant stride in machine learning, offering the capability to generate coherent natural language text based on given contexts (Shanahan, 2023). The advent of InstructGPT (Ouyang et al., 2022) epitomizes this progression, heralding a new era of LLMs with enhanced instruction-following and logical reasoning skills. Although proprietary models like OpenAI's GPT-3.5 and GPT-4 set performance benchmarks, the rise of open-source LLMs presents a compelling narrative of achieving comparable state-of-the-art (SOTA)

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 305–310
June 20-21, 2024 ©2024 Association for Computational Linguistics

performance with cost-effective implementations (Li et al., 2023)(Jiang et al., 2024).

Multi-Agent Collaboration Drawing parallels to human teamwork, integrating LLMs as collaborative agents has shown improved efficacy across diverse tasks. Initiatives like BabyAGI (Nakajima, 2023) introduced frameworks for automatic task generation and execution, based on predefined objectives. AutoGPT (aut, 2023) extends LLMs' capabilities to interact with external tools for executing real-world tasks, such as web scraping and code execution. Furthermore, HuggingGPT (Shen et al., 2023) functions as a model selector within the Hugging Face ecosystem, optimizing task-specific model selection. MetaGPT (Hong et al., 2023) emulates a software development team, assigning distinct roles to LLMs to streamline the design and development process. This body of work underscores the significant enhancements and novel functionalities afforded by multi-agent collaboration.

LLM Debating System Debates, a cornerstone in assessing the viability of ideas within human discourse, have been adapted to the realm of LLMs. Initial investigations by (Liang et al., 2023) into multi-agent debating revealed that a structured, mildly antagonistic debate could refine LLM outputs. Subsequent research (Xiong et al., 2023) corroborated the potential of LLMs to achieve consensus through debate. However, studies by (Chen et al., 2023) and (Agashe et al., 2023) on the evaluation of multi-agent debating systems highlighted a critical issue: the risk of consensus being swayed by majority opinion rather than individual agent analysis. This introduces an element of uncertainty regarding whether the consensus reached is genuinely reflective of a reasoned agreement or merely a product of majority rule. This paper seeks to explore and address this ambiguity in the context of LLM debating systems.

2.2 Dataset Discription

The dataset for our study, meticulously curated by clinical domain experts, trial organizers, and research oncologists affiliated with the Cancer Research UK Manchester Institute and the Digital Experimental Cancer Medicine Team (Jullien et al., 2023), comprises the following elements:

- **1–2 CTRs**: Record some key information during clinical trial, constitute by these four parts:
  - **Eligibility Criteria:** Specifies the required conditions for patients to participate in the clinical trial.
  - **Intervention Details:** Outlines the type, dosage, frequency, and duration of the treatments under study.
  - **Trial Results:** Details the number of participants, outcome measures, measurement units, and the observed results.
  - **Adverse Events Reporting:** Records any symptoms or signs noted in patients during the course of the clinical trial.

- **Statement**: An assumption based on CTRs, which hasn't been verified to be correct or not

- **Section Marker**: Which section in the CTRs is the statement based on.

- **Entailment/Contradiction label**: The statement is Entailment/Contradiction to the CTRs.

Table 1 describe the constitute of the dataset, and Table 2 is a example of test data.

Dataset	Comparison	Single	Total
Train	665	1035	1700
Dev	60	140	200
Test	2947	2553	5500

Table 1: Constitute of the dataset.

Attribute	Value
Type	Single
Section_id	Results
Primary_id	NCT02640053
CTR_context	Outcome Measurement: Area Under the Curve (AUC) EORTC CIPN20 Sensory Neuropathy Subscale...(omitted)
Statement	Patients in the primary trial that didn't receive topical cryotherapy had worse symptoms than patients that did receive topical cryotherapy.
Label	Contradiction

Table 2: Test data example.

3 System Overview

3.1 Motivation

Existing multi-agent collaboration frameworks, while adept at executing tasks like coding, often fall short in fostering substantive dialogues among agents. This limitation hinders the development of critical thinking skills, as agents are not encouraged to engage in detailed discussions or critically evaluate one another's viewpoints. Recognizing this gap, we introduce a novel framework designed specifically to enable multi-agent debate. Our approach centers on facilitating a collaborative environment where agents are encouraged to thoroughly consider and reflect on the perspectives of their counterparts. By prioritizing in-depth discussions and critical analysis, we aim to advance the capabilities of multi-agent systems beyond mere task execution to include nuanced, critical deliberations.

3.2 Multi-Agent Debating Framework

The multi-agent debating framework constitute by several costume *agents*, a *issue* to determine and a *logical judgment unit*. Below Algorithm 1 are pseudocode that describe how the framework operates:

Algorithm 1: LLM Multi-Agent Debate Framework

Data: *Issue* to be debated, *Agents* involved in the debate
Result: Conclusive outcome(**Entailment** or **Contradiction**)

```
initialization: Set turn t_i = 0;
Agents generate initial responses r_i with Opinion and Decision;
while not reached maximum number of turns and no consensus do
    Assess consensus among Opinions;
    if consensus then
        Adopt Opinions as outcome and terminate;
    else
        Increment turn t_{i+1};
        Update agents with others' Decisions;
        Agents revise responses r_{i+1};
    end
end
if no consensus after maximum turns then
    Take most Opinions as final result;
end
```

Upon presenting an issue, the framework, in its initial turn denoted as t_i, solicits from agents the generation of an initial response r_i. Each response is required to concurrently encompass ***Opinion*** and ***Decision***, wherein ***Opinion*** constitutes a paragraph articulating the agent's stance on the issue, and ***Decision*** represents one of two potential outcomes: Entailment or Contradiction. Subsequent to the formulation of responses by all agents, the logical judgment unit assesses the presence of consensus within their opinions. In the event of consensus, their ***Opinions*** are adopted as the conclusive outcome, thereby terminating the framework. Conversely, in the absence of consensus, the debate advances to the subsequent turn t_{i+1}. In this phase, each agent is apprised of the ***Decisions*** made previously by other agents and is prompted to generate a revised response r_i. This iterative process persists until a consensus is established among the agents, or upon reaching the predefined maximum number of turns, at which point the framework is concluded and the amalgamation of multiple ***Opinions*** is deemed the final result.

4 Experiments

Our experiments were conducted using the Mixtral-8x7B model (Jiang et al., 2024), selected for its exceptional performance and cost-efficiency.

4.1 Sections Select

The task of pinpointing the relevant sections within Clinical Trial Reports (CTRs) for statement verification was entrusted to a Large Language Model (LLM). This procedure entailed providing the LLM with a detailed prompt, encompassing explicit instructions, the statement under scrutiny, and the entirety of the CTR text. The LLM's assignment was to ascertain the sections of the CTR pertinent to the statement. An example of the LLM's output is delineated below, illustrating its capability to effectively identify and isolate relevant text segments.

Listing 1: LLM's output to select sections

```
{
  "Primary_CT": {
    "Adverse Events": true,
    "Results": false,
    "Eligibility": true,
    "Intervention": false
  }
}
```

4.2 Agents Design

We designed five agents, which are all experts in medical field:

- **Dr. Emily Nguyen**: Biostatistician focusing on data interpretation and analysis in clinical trials.
- **Dr. Alex Johnson**: Medical Linguist specializing in clinical text analysis and medical jargon clarification.
- **Dr. Aisha Patel**: Pharmacologist dedicated to drug action understanding and safety evaluation in trials.
- **Dr. Liang Wei**: Epidemiologist studying health and disease patterns in populations for disease control.
- **Dr. Maria Gomez**: Cardiologist treating cardiovascular diseases and managing heart-related conditions.

5 Results

5.1 Official Evaluation Metrics

SemEval-2024 Task 2 organizers had mentioned several evaluation metrics: Macro F1-score, Faithfulness and Consistency(Jullien et al., 2024a), we will use these metrics to evaluate the result.

- **Faithfulness**: Quantifies the precision with which a system arrives at the correct conclusion based on the right reasons. Assessed by examining the system's ability to adjust its predictions accurately in response to semantic alterations. Evaluated using N statements x_i from a contrast set (C), their related original statements y_i, and the model's predictions $f()$.
- **Consistency**: Measures a system's capacity to produce identical outputs for semantically equivalent scenarios. Determined by the system's capability to consistently predict the same label for both original and contrast statements, even in the case of semantically preserving interventions. This involves analyzing N statements x_i in a contrast set (C), their corresponding original statements y_i, and the model's predictions $f()$.

Agent	F1	Fth	Con
Debated	**0.69**	0.51	0.56
Dr. Emily Nguyen	0.65	**0.55**	**0.58**
Dr. Alex Johnson	0.67	0.52	0.56
Dr. Aisha Patel	0.67	**0.55**	0.57
Dr. Liang Wei	0.66	0.53	0.57
Dr. Maria Gomez	0.66	0.53	0.57

F1 = Macro-F1, Fth = Faithfulness, Con = Consistency

Table 3: Performance metrics of different agents.

5.2 Impact of Multi-Agent Debates on Performance

Table 3 delineates the comparative performance metrics on the test dataset for decisions made by individual agents versus the consolidated outcome of the multi-agent debating, as evaluated on the official Codalab platform (Jullien et al., 2024b). The data distinctly showcases that the integration of multi-agent debating significantly enhances the F1-score, surpassing the performance of any single agent's decision-making capabilities.

Regarding our standing in the competition, our team, identified as "Yuan_Lu," achieved the 22nd position out of 32 participating teams.

5.3 Demonstration of Debating Process

The excerpt below showcases the dynamic of our multi-agent debating process, wherein agents engage in a layered analysis of the problem from their unique vantage points. Notably, our methodology facilitates not just the articulation of each agent's perspective but also the receptiveness to alternate viewpoints. This iterative exchange allows agents to both influence and be influenced, showcasing the framework's capacity for fostering a comprehensive and nuanced understanding through collaborative debate.

5.3.1 Turn 1

DR. EMILY NGUYEN: Based on the data provided, I cannot confirm the statement about the occurrence of asthenia and pyrexia... **The data does not include any information on these adverse events** for either the primary or secondary clinical trials... So, my decision is: **Contradiction**

DR. ALEX JOHNSON: After carefully reviewing the data provided and considering the opinions expressed, I agree that **there is**

no evidence in the data to support the statement about the occurrence of asthenia and pyrexia... So, my decision is: **Contradiction**

5.3.2 Turn 2

DR. EMILY NGUYEN: After further discussion and considering the perspectives shared, **I now see the importance of rigorous data analysis and transparent reporting**. This ensures healthcare professionals and patients have a complete understanding... In summary, **the provided data does not support the statement**... So, my decision is: **Contradiction**

DR. ALEX JOHNSON: Reflecting on the discussions, I acknowledge my initial decision might have been premature. **Given the lack of evidence in the provided data**, I adjust my stance... Therefore, **I align with the consensus that the statement is a contradiction**... So, my decision is: **Contradiction**

5.4 Beyond Majority Rule: The Dynamics of Multi-Agent Debating

To examine the effectiveness of the multi-agent debating framework, we introduce two key metrics: the F1 score of the majority decision in the first round of debate, and the rate of alignment between the initial majority decision and the final decision.

- **F1 Score of Correct Majority Decision**: This metric measures the precision and recall of the majority decision in the first round of debate. The calculated F1 score of 0.567 suggests moderate accuracy of the initial majority opinion. This indicates that while the majority decision is somewhat aligned with the correct labels, there is still considerable room for improvement.

- **Rate of Alignment Between First Majority and Final Decision**: This metric reflects the consistency between the initial majority decision and the final decision after the debate. A high rate of 91.93% suggests that the final decisions of the multi-agent system are predominantly in line with the initial majority opinions. This high alignment rate might indicate a strong influence of the initial majority view on the final outcome of the debate.

Considering these results, the multi-agent debating framework demonstrates a significant tendency to converge towards the initial majority decision. However, the fact that the final decision's F1 score is 0.69, which is higher than the initial majority's F1 score, indicates that the debating process adds value beyond simply following the majority rule. This suggests that while the final decision often aligns with the initial majority opinion, the debate process itself contributes to refining the decision, potentially correcting or enhancing the initial judgment. Therefore, despite the high alignment rate, the multi-agent debating framework plays a critical role in facilitating a more comprehensive and informed decision-making process.

6 Conclusion

We have presented a novel multi-agent debating framework to participate SemEval-2024 Task 2. This approach, integrating the expertise of diverse agents like Biostatisticians, Medical Linguists, and Pharmacologists, significantly enhances the analysis of Clinical Trial Reports (CTRs). Our findings demonstrate improved performance in entailment or contradiction determination of CTR-statement pairs, as evidenced by enhanced Macro F1-scores compared to individual agent assessments. Despite a tendency to align with initial majority decisions, the debating process refines these initial judgments, indicating the framework's effectiveness beyond simple majority rule.

References

2023. Auto-gpt.documentation.

Saaket Agashe, Yue Fan, and Xin Eric Wang. 2023. Evaluating multi-agent coordination abilities in large language models.

F. Amar, A. April, and A. Abran. 2024. Electronic health record and semantic issues using fast healthcare interoperability resources: Systematic mapping review. *Journal of Medical Internet Research*, 26:e45209.

Huaben Chen, Wenkang Ji, Lufeng Xu, and Shiyu Zhao. 2023. Multi-agent consensus seeking via large language models.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2023. Metagpt: Meta programming for a multi-agent collaborative framework.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

Mael Jullien, Marco Valentino, and André Freitas. 2024a. Semeval-2024 task 2: Safe biomedical natural language inference for clinical trials. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1948–1963, Mexico City, Mexico. Association for Computational Linguistics.

Mael Jullien, Marco Valentino, Hannah Frost, Paul O'Regan, Dónal Landers, and Andre Freitas. 2023. NLI4CT: Multi-evidence natural language inference for clinical trial reports. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16745–16764, Singapore. Association for Computational Linguistics.

Maël Jullien, Marco Valentino, and André Freitas. 2024b. CodaLab - Competition — codalab.lisn.upsaclay.fr. `https://codalab.lisn.upsaclay.fr/competitions/16190`. [Accessed 13-02-2024].

Shiyang Li, Jun Yan, Hai Wang, Zheng Tang, Xiang Ren, Vijay Srinivasan, and Hongxia Jin. 2023. Instruction-following evaluation through verbalizer manipulation.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging divergent thinking in large language models through multi-agent debate.

Yohei Nakajima. 2023. Babyagi. GitHub repository.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

M. Saban, M. Lutski, I. Zucker, M. Uziel, D. Ben-Moshe, A. Israel, S. Vinker, A. Golan-Cohen, I. Laufer, I. Green, R. Eldor, and E. Merzon. 2024. Identifying diabetes related-complications in a real-world free-text electronic medical records in hebrew using natural language processing techniques. *Journal of diabetes science and technology*, page 19322968241228555.

Murray Shanahan. 2023. Talking about large language models.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face.

Kai Xiong, Xiao Ding, Yixin Cao, Ting Liu, and Bing Qin. 2023. Examining inter-consistency of large language models collaboration: An in-depth analysis via debate.