Interruptible Multi-Agent Debate: Sentence-Level Disclosure and Urgency-Based Turn-Taking for Early Error Correction

Akikazu Kimura[1], Ken Fukuda[2], Yasuyuki Tahara[3], Yuichi Sei[4]*
The University of Electro-Communications, Tokyo, Japan[1,3,4]
National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan[1,2]

***Abstract*—Multi-agent debate (MAD) has emerged as a promising approach for improving the reasoning ability of large language models (LLMs). However, existing turn-taking schemes typically disclose a speaker's entire utterance before other agents can respond, allowing erroneous premises to spread through the shared context and making early correction difficult. This study proposes an interruptible MAD framework that enables early error correction through sentence-level disclosure and urgency-based turn-taking under a shared public-token budget. Each non-speaking agent continuously generates an action plan including its current assessment, action choice, urgency, and supported answer. The next speaker is then selected dynamically from agents requesting to speak or interrupt, while silent turns are allowed when no intervention is necessary. By revealing only one sentence at a time and discarding undisclosed sentences after interruption, the proposed framework is designed to prevent misleading claims from expanding into long incorrect explanations. Under a controlled evaluation on 1,000 MMLU questions using three agents with conditioned initial states containing both correct and incorrect answers, the proposed framework achieves the highest final accuracy in both the two-incorrect-one-correct setting (49.5% vs. 37.2% and 43.7%) and the one-incorrect-two-correct setting (79.2% vs. 68.7% and 73.8%). Analysis of intermediate answers further show that interruptions improve listeners' answers more often than they worsen them. These results suggest that fine-grained, interruptible turn-taking can suppress misinformation propagation and stabilize consensus formation under the evaluated setting.**

***Keywords*—Large Language Models (LLMs); multi-agent debate; turn-taking; early error correction; misinformation propagation**

I. Introduction

With the rapid development of large language models (LLMs), multi-agent debate (MAD), in which multiple LLM agents reach consensus by mutually criticizing and verifying one another, has attracted growing attention for tasks that require complex reasoning and decision-making. MAD has been reported to improve reasoning ability and factuality compared with a single LLM [1]. At the same time, MAD performance depends not only on the capabilities of individual agents but also strongly on the turn design that controls how the discussion unfolds.

Many existing MAD frameworks rely on fixed speaking orders or rule-based speaker transitions, and speakers generally take turns in a predetermined sequence. In such fixed-order turn-taking, even when the beginning of an utterance contains an incorrect premise, the other agents cannot speak until their turns arrive. As a result, they cannot intervene at arbitrary moments in the way humans often do in conversation.

To address this limitation, prior work has proposed methods in which agents bid for the right to speak and the next speaker is selected according to the bid [2]. However, once a speaker has obtained the floor, other agents still cannot intervene until that speaker has finished the utterance. Consequently, these methods cannot prevent a long explanation from being generated on the basis of an incorrect premise introduced at the beginning of the utterance. This limitation is problematic from the viewpoint of early error correction.

LLMs are known to be susceptible to errors in the input or immediate context, and misinformation can propagate through the reasoning process. It has also been suggested that correcting errors earlier is more effective [3]. Therefore, when an utterance begins with an incorrect premise, a long line of reasoning built on that premise can dominate the shared context, encourage other agents to follow it, and amplify misinformation. Indeed, previous work has reported that, in MAD, introducing adversarial agents that intentionally steer the discussion toward wrong answers degrades problem-solving accuracy [4].

In conversation analysis (CA), interruptions and silence at turn-transition points function as resources for adjusting floor ownership and prompting subsequent speech [5]. Moreover, when errors or inconsistencies arise during conversation, earlier initiation of repair suppresses the spread of later misunderstandings more effectively [6]. These observations suggest that MAD also requires flexible turn-taking in which listeners remain silent while trying to understand a claim, but interrupt immediately when they detect an error that should be corrected.

Despite recent progress in dynamic speaker selection, a key structural limitation remains in existing MAD frameworks: the unit of public disclosure is still typically a full message. As a result, once a speaker obtains the floor, listeners must often wait until the end of the utterance even if an incorrect premise appears in its opening sentence. This design delays repair and allows misleading reasoning to accumulate in the shared context.

Accordingly, the goal of this study is not to claim universal improvements across all models and tasks, but to examine under controlled conditions the extent to which finer-grained

*Corresponding author.

disclosure and interruptible turn-taking can reduce the spread of incorrect premises during debate. In particular, we focus on the debate-control mechanism itself and investigate whether allowing agents to react at the sentence level enables earlier repair before an incorrect line of reasoning expands into a long public explanation.

Motivated by this, we propose a MAD framework that introduces a turn-taking mechanism in which utterances are disclosed sentence by sentence and the next action is determined dynamically according to the urgency of the responses generated by listeners. By explicitly incorporating interruption and silence, the proposed method allows agents to intervene as soon as an error is detected and begin repair immediately. In this way, it aims to improve accuracy on reasoning tasks with a single correct answer while avoiding long explanations grounded in incorrect premises and minimizing the propagation of misinformation.

This study does not assume that a discussion can continue indefinitely. Instead, we consider a setting in which consensus must be formed under limited shared resources. Rather than controlling discussion length by the number of utterances, we use the cumulative number of publicly disclosed utterance tokens. Public utterances constitute the shared context and impose an external cost because they consume the reasoning and attention resources of the other agents. In contrast, the computational resources consumed internally by each agent are private resources and do not directly burden others. Accordingly, we treat public utterance tokens as the shared resource that should be constrained in the discussion. Furthermore, if agents represent different people or organizations, or if humans and agents participate together, the effective constraint is not internal computation but the amount of information presented to others. From this perspective as well, controlling discussion by the total amount of public utterance tokens and the cumulative number of disclosed tokens is consistent with real-world settings.

This study makes three main contributions. First, it proposes an interruptible multi-agent debate framework that combines sentence-level disclosure, urgency-based speaker selection, interruption, and silent turns under a shared public-token budget. Second, it introduces a controlled evaluation setting in which initial answers are aligned across conditions so that differences can be attributed more directly to debate control rather than to initial sampling variance. Third, it provides not only final-answer comparisons but also process-level analysis of intermediate answers, interruption events, and completion events to examine how misinformation is corrected or propagated during debate.

The remainder of this study is organized as follows. Section II reviews related work. Section III describes the proposed framework in detail. Section IV presents the experimental settings, and Section V reports the main results. Section VI analyzes and discusses the effectiveness of the proposed framework. Finally, Section VII concludes the study and outlines future directions.

II. Related Work

A. Multi-LLM Agent Debate

Research on coordinating multiple LLMs as agents has developed as a way to improve performance on complex tasks. For example, CAMEL, AutoGen, and ChatDev demonstrated frameworks in which role assignment and dialogue enable agents to solve tasks that are difficult for a single agent to handle [7], [8], [9]. Building on this line of work, multi-agent debate (MAD) has been reported to improve reasoning ability and factuality [1], as well as evaluation performance [10], by having multiple agents verify one another's claims. It has also been shown that diversity among agents and the introduction of reflection can improve discussion quality [11]. However, in most existing MAD frameworks, the unit of speaker change is an entire message, and utterances are not disclosed sentence by sentence. As a result, other agents cannot interrupt while a speaker is still in the middle of an utterance.

B. Dynamic Speaker Selection

In parallel, recent work has explored the introduction of human-like flexibility in speaker selection. In Werewolf Arena, Bailis et al. proposed a method in which each non-speaking LLM agent listens to the ongoing utterance and outputs, at every turn, a bid representing its desire to speak next; the next speaker is then selected according to that value [2]. Similarly, prior studies have improved the naturalness and coherence of multi-agent conversations by selecting the next speaker based on importance scores output by non-speakers and by introducing concepts such as adjacency pairs and turn-taking from conversation analysis into MAD [12], [5]. These approaches make it possible to approximate more autonomous, human-like discussion, for example by assigning high scores when an utterance is directed to a particular agent. Nevertheless, in all of these studies, the disclosure unit remains the entire generated utterance. Therefore, if an incorrect premise appears at the beginning of an utterance, other agents still cannot interrupt and correct it at that moment.

C. Misinformation Propagation and Early Repair

LLMs are vulnerable to errors in the input and immediate context, and misinformation can spread recursively through the reasoning process [3]. Feng et al. showed that when an LLM is given an incorrect premise as input, misinformation propagates through subsequent reasoning and substantially reduces the final answer accuracy. They also showed that earlier correction is more effective, suggesting that beginning repair immediately after an incorrect premise is introduced is important for minimizing misinformation propagation [3]. Moreover, even in MAD settings, introducing adversarial agents that intentionally steer the discussion toward wrong answers causes incorrect claims to spread to other agents through the shared context and degrades reasoning performance [4]. These findings indicate that MAD requires a flexible turn-taking mechanism that allows listeners to interrupt immediately and initiate repair when a speaker presents an incorrect premise.

III. Proposed Framework

A. Framework Overview

We propose a multi-agent debate framework designed to prevent discussions from developing on the basis of incorrect premises while enabling the flexible transfer of floor ownership observed in human conversation.

In this framework, multiple LLM agents discuss a given task and derive a final answer. This section describes the three stages that constitute the framework: 1) generation of initial answers, 2) multi-turn debate, and 3) extraction of the final answer.

B. Initial Answer Generation

At this stage, each agent generates an initial answer to the given question together with the reason for choosing it (`answer`, `reason`). To obtain answer choices and reasons in a consistent format, the initial answer is generated with a prompt that encourages chain-of-thought (CoT) reasoning. CoT has been reported to improve performance on multi-step reasoning problems by prompting large language models to describe intermediate reasoning steps [13]. In the multi-turn debate stage, these answers and reasons are embedded in the prompt and shared among the agents.

C. Multi-turn Debate

In the multi-turn debate stage, each agent discusses the problem on the basis of the initial answers and collectively works toward a final answer. Below, we describe the components introduced to realize the flexible transfer of speaking rights that is characteristic of human discussion.

1) Action plan generation: In the proposed framework, only publicly disclosed utterances are counted as shared resources, and the total amount of such utterances shared across the debate is managed as `token_budget`. In addition, the total amount of utterances disclosed up to time t is referred to as the cumulative number of public utterance tokens. At each turn, the remaining number of public utterance tokens, `tokens_left`, is computed from this value and presented to all agents as a common input when generating action plans. During the debate, each non-speaking agent receives, as the current debate state, the information listed in Table I.

TABLE I. Debate Information Received by Each Agent

Variable	Description
token_budget	Total public utterance token budget shared across the entire debate
tokens_left	Remaining public utterance tokens currently available
initial_answer	Initial answers of all agents and their reasons
turn_log	Log of utterances up to the previous turn
previous_thoughts	Log of previous thoughts
last_event	Speaker's utterance in the current turn

Based on this debate information, each agent outputs an action plan in JSON format. The output fields are shown in Table II. When determining the speaker for the first turn or for a turn in which no current speaker exists, `Action` can be either `listen` or `speak`. When there is an active speaker, `Action` can be either `listen` or `interrupt`.

TABLE II. Output Format of the Action Plan

Field	Description
Thought	Reflective thought about the current debate state
Action	Desired action for the next turn (`listen`, `speak`, or `interrupt`)
Urgency	Urgency score for speaking in the next turn (0–9)
Purpose	Purpose of the selected action
Answer	Answer currently supported most strongly

The prompt used for action-plan generation is shown in Algorithm 1. This prompt takes the debate information listed in Table I as input and outputs a JSON object consisting of `thought`, `urgency`, `action`, `purpose`, and `answer`. Among these fields, `action` and `urgency` are used to determine the next speaker, whereas `answer` is recorded as the intermediate answer at each turn. The available choices for `action` depend on the state of the debate. If a current speaker exists, the agent selects either `listen`, meaning that it continues listening to the current utterance, or `interrupt`, meaning that it intervenes immediately. If no current speaker exists, such as at the beginning of the debate or after a silent turn, the agent selects either `listen` or `speak`, the latter meaning that it starts a new utterance. The prompt also asks the agent to output a brief evaluation of the current-turn utterance as `thought` and then judge, as `urgency`, how necessary it is to speak immediately. In particular, the prompt instructs the agent to assign high urgency when it can correct a factual or logical error in the latest utterance, when few public utterance tokens remain, or when its current answer disagrees with the apparent majority in the debate history. By contrast, if the utterance may still be in progress and its continuation may resolve the concern, the agent is encouraged to choose `listen`, thereby suppressing unnecessary interruptions. To facilitate reproducibility, we make the implementation of the proposed framework publicly available at https://github.com/KimuraAkikazu/dynamicdebate. The repository includes the prompt templates used in the experiments, together with the main execution scripts, configuration files, and analysis scripts.

In the proposed framework, urgency is used as an operational coordination signal rather than as a calibrated probability of correctness. The score is used only for relative ranking among agents at a given turn to determine who should speak next. This design enables decentralized speaker selection without introducing an additional controller model or handcrafted interruption rules, while keeping the mechanism lightweight and reproducible. At the same time, the present study does not claim that self-reported urgency is perfectly calibrated across tasks or models. Its calibration should therefore be interpreted as a limitation of the current framework and as an important direction for future work.

2) Speaker selection for the next turn: The next speaker is determined on the basis of the action plans of all agents. Candidates are limited to the agents that selected `speak` or `interrupt` as `Action`, and the agent with the highest `urgency` is selected as the next speaker. If multiple agents tie for the highest urgency, one of them is chosen at random so as to avoid bias from a fixed ordering.

If no candidate exists, the next turn becomes a silent turn.

Algorithm 1: Action Plan Prompt

```
You are {name}. You are debating to arrive at the
    correct answer to the question. Based on the
    debate information, generate a response to the
    instructions.

## Question
{topic}
# Debate Context
## Debate rules
- The debate ends when the shared public token
    budget of {token_budget} tokens is exhausted.
    Only revealed text counts toward this budget.
- Remaining public tokens available to all agents:
    {tokens_left} / {token_budget}.
- Only one member can speak per turn. The next
    speaker is selected from the highest urgency
    level.
- Each turn, one chunk at a time from the speaker's
    generated statement is revealed to all members.
    The current speaker may have more statement
    prepared and not yet revealed.

## Initial answers before debate begins (subject to
    change)
{initial_answer}

## Debate history
{turn_log}

## Your memory
{previous_thoughts}

## Event of this turn
{last_event}

## State
- This is turn {turn}. Use the remaining public
    tokens efficiently to reach the correct answer.

# Actions
You can take the following actions:
- 'listen': Use this when you listen to the
    speaker's argument until the end to deepen your
    understanding before speaking.
- 'interrupt': Use this when you interrupt the
    current speaker to begin speaking.

# Instructions
1. Based on the debate so far and the utterance of
    this turn, briefly explain your current
    internal thoughts such as your perspective on
    the responses to the questions, your action
    plan for the remaining turns, your reaction.
2. Determine the urgency for you to start talking
    now. Raise urgency only when 1) you can
    immediately correct a factual or logical error
    in the latest statement, 2) tokens_left is low
    and essential information must be shared soon,
    or 3) your current answer disagrees with the
    apparent majority in the debate history.
    Otherwise keep urgency low and listen.
3. Refer to the provided information and your
    current thought, decide the next turn's action
    you should take as {name}.
- While considering the possibility that someone
    may be mid-sentence, decide whether to
    interrupt and respond immediately to this
    turn's statement or listen to its completion.
4. Select the purpose of the action you have chosen.
5. Based on the debate information, output the
    answer to the question that you currently
    support the most.
- If the anticipated continuation of statement may
    resolve your concern, choose listen.
```

In a silent turn, no utterance is produced, and all agents update their action plans for the following turn.

3) Utterance generation and sentence-by-sentence disclosure: The agent selected as the next speaker generates an utterance on the basis of the debate history and its own action plan. The generated utterance is segmented into sentences and stored in an internal queue. Sentence segmentation is performed using the SpaCy natural language processing library*. We adopt sentences rather than fixed-length token chunks because sentences preserve semantic units more naturally and are closer to the minimal unit of human conversational contribution. During the debate, only one sentence is disclosed from the queue in each turn. This design makes it easier for other agents to intervene in the immediately following turn, even when an incorrect premise appears at the beginning of an utterance.

The disclosed sentences are accumulated as debate history and are then used as input for subsequent action-plan generation and utterance generation.

Sentence segmentation was adopted because sentences preserve semantic units more naturally than fixed-size token chunks and provide a practical unit for potential interruption. In this study, segmentation is performed deterministically with SpaCy for English debate outputs. However, segmentation quality may affect interruption timing, especially when generated text contains ellipsis, coordination, or non-canonical punctuation. The reported results should therefore be interpreted as evidence for the proposed mechanism under this segmentation setting rather than as a claim of segmentation invariance.

4) Definition of interruption: We define an interruption as a speaker change that occurs while a current speaker still has undisclosed sentences remaining in its queue and another agent is selected as the next speaker. When an interruption occurs, the next speaker begins speaking immediately in direct response to the preceding utterance. At that point, the original speaker's undisclosed sentences are discarded. By allowing this type of interruption, the proposed framework makes it easier for correction to occur before an explanation based on an incorrect premise expands over multiple sentences.

D. Final Answer Extraction

At each turn, the number of tokens in the disclosed utterance is counted incrementally and recorded as the cumulative number of public utterance tokens. The debate ends when this cumulative count reaches a predefined threshold. At the turn in which the threshold is reached, the `Answer` fields contained in the agents' action plans are collected as final-answer candidates, and the majority answer is adopted as the final answer.

IV. Experimental Settings

As the number of agents and the length of the discussion increase in multi-agent debate, more reasoning and rebuttal become possible, but the computational cost and total token consumption also grow rapidly, making it difficult to secure a large experimental scale and to compare conditions fairly. If debates are allowed to continue indefinitely, evaluation

*https://spacy.io/

becomes difficult not only because of computational constraints but also because the amount of discussion can differ substantially across conditions. For this reason, existing studies commonly evaluate under settings with upper bounds on the number of utterances or the number of agents [1], [14]. Similarly, prior work on dynamic turn-taking [2], [12] imposes explicit limits on game progression or discussion length in order to make evaluation feasible.

Based on these considerations, we set the upper limit of the total public utterance token budget to 500, which allowed sufficient room for discussion to develop while keeping execution costs manageable.

Table III summarizes the debate settings used in this study.

TABLE III. DEBATE SETTINGS USED IN THE EXPERIMENTS

Item	Value
Total Public Utterance Token Budget	500 tokens
Number of agents	3

The experiments in this study are designed as a controlled mechanism study rather than as a broad generalization study. A single model and a single benchmark are used so that differences among conditions can be attributed as directly as possible to turn-taking control. Likewise, conditioned initial states are used to reproduce situations in which correct and incorrect answers coexist and to isolate how each framework responds to misinformation already present in the debate. The results therefore establish the effectiveness of the proposed design under the evaluated setting, while broader validation across models, domains, and agent counts is left for future work.

A. LLM

We used Meta's Llama-3.1-8B-Instruct quantized to 8-bit[†]. The model parameters used in the experiments are listed in Table IV.

TABLE IV. LLM PARAMETERS USED IN THE EXPERIMENTS

Model	Llama-3.1-8B-Instruct
Context length	15000 tokens
Maximum tokens	1024 tokens
Temperature	0.3

B. Tasks and Dataset

To evaluate the effectiveness of the proposed framework, we adopted Massive Multitask Language Understanding (MMLU), which consists of multiple-choice questions across diverse subjects centered mainly on high-school-level knowledge [15]. MMLU covers a wide range of domains, including STEM, the humanities, and the social sciences, and has been widely used as a benchmark for evaluating the general knowledge and reasoning abilities of language models [16], [17], [18].

To isolate the effect of debate control, all compared conditions started from the same initial answers.

[†]https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

In addition, to examine whether the introduction of interruption and silence can suppress the propagation of misinformation, we conditioned the initial answers. Specifically, the discussion began from a state in which a particular agent already held misinformation, allowing us to evaluate how each condition was affected by misinformation propagation.

More concretely, for each question randomly sampled from MMLU, we generated answers five times with a single LLM and stored the answer and reason obtained in each run. We retained only those questions for which both correct and incorrect answers appeared at least twice among the five outputs. At debate start, three of the five stored (answer, reason) pairs were assigned to the three agents so as to realize the desired configurations (two incorrect and one correct, or one incorrect and two correct), and these were embedded into the prompt as the agents' initial answers. This procedure allowed us to align the initial state—that is, the initial answers and reasons of all agents—across conditions and evaluate only the difference in debate control.

C. Compared Conditions

Rather than reproducing individual prior methods exactly, we compared abstracted conditions that capture representative turn designs in the literature so that the effects of speaking order and disclosure granularity could be disentangled. Specifically, the fixed-order condition abstracts fixed-turn MAD [1], [10], whereas the dynamic-order condition abstracts methods with dynamic speaker selection [2], [12]. To isolate the effect of the proposed debate control with interruption and silence, we compared the following three conditions.

1) Fixed order: Speakers alternated in a predetermined order, and each utterance was disclosed as a full generated message.

2) Dynamic order: At each turn, non-speakers declared their desire to speak based on urgency, and the next speaker was selected accordingly. However, utterances were disclosed at the message level, and speaker changes did not occur until the current speaker had finished.

3) Proposed framework: In addition to urgency-based dynamic speaker selection, the framework introduced sentence-by-sentence disclosure, urgency-based interruption, and silent turns, as described in Section III.

D. Evaluation Metrics

In each condition, debates were conducted on 1,000 questions from the constructed dataset. As evaluation metrics, we used the intermediate answers output by each agent in its action plan at each turn, evaluated the change in accuracy over turns, and thereby assessed the susceptibility of each method to misinformation. We also evaluated the accuracy of the final answers at the end of the debate to determine whether the proposed framework could improve accuracy while suppressing the propagation of misinformation. For reproducibility, the implementation, major prompts, evaluation scripts, and execution settings are publicly available at https://github.com/KimuraAkikazu/dynamicdebate.

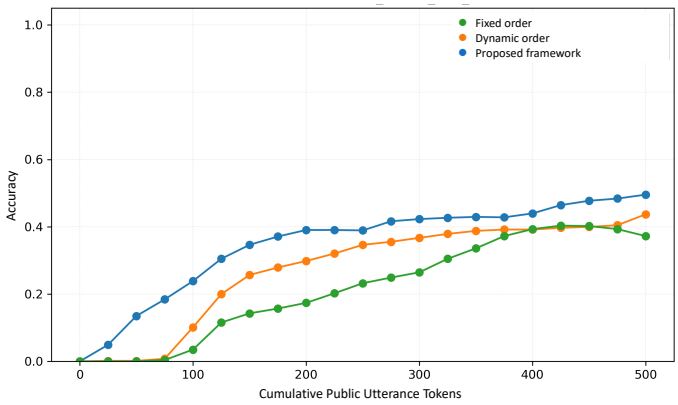

Fig. 1. Trajectory of intermediate-answer accuracy under each condition when two agents were initially incorrect.

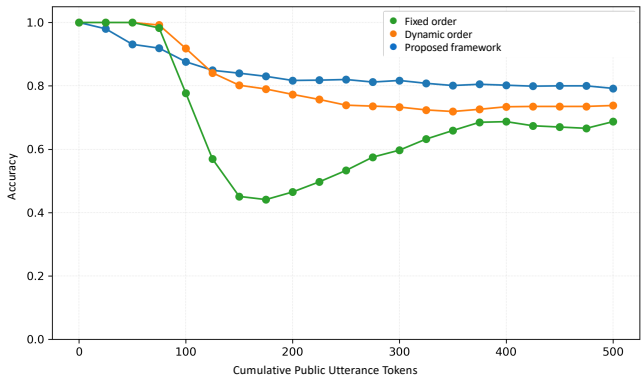

Fig. 2. Trajectory of intermediate-answer accuracy under each condition when one agent was initially incorrect.

V. Experimental Results

To examine the effectiveness of the proposed framework, we conducted evaluation experiments under the settings described in Section IV. In this section, we summarize the effects of the proposed framework from the viewpoints of final task accuracy, the trajectory of accuracy during the discussion, and answer changes caused by interruptions.

A. Accuracy

This subsection evaluates how the presence of agents holding misinformation affected debate accuracy. We compared final accuracy and the trajectory of intermediate answers under two settings that differed in the number of initially incorrect agents.

When two agents were initially incorrect, the proposed framework achieved 49.5%, which was 12.3 points higher than Fixed order (37.2%) and 5.8 points higher than Dynamic order (43.7%), as shown in Fig. 1 and Table V. Compared with Fixed order, the two dynamic conditions improved accuracy in the early stage of discussion and showed a more stable trajectory as the debate progressed. This early-stage gain was especially pronounced for the proposed framework.

TABLE V. Final Accuracy Under Each Condition

Condition	Accuracy
Fixed order	37.2%
Dynamic order	43.7%
Proposed framework	49.5%

These results indicate that the proposed framework is especially beneficial when the debate starts from an incorrect majority, suggesting that early interruption helps the correct side intervene before the incorrect majority becomes entrenched in the shared context.

When only one agent was initially incorrect, the proposed framework again achieved the highest accuracy at 79.2%, which was 10.5 points higher than Fixed order (68.7%) and 5.4 points higher than Dynamic order (73.8%), as shown in Fig. 2 and Table VI. In the two conditions without interruption, accuracy tended to drop sharply in the early stage. By contrast, in the proposed framework, the initial drop was comparatively suppressed and accuracy remained more stable as the debate progressed. In particular, under Fixed order, the incorrect agent was scheduled to speak first, causing a substantial initial drop in accuracy. Although some recovery was observed in the later stage, the final accuracy remained the lowest among the three conditions.

TABLE VI. Final Accuracy Under Each Condition When One Agent was Initially Incorrect

Condition	Accuracy
Fixed order	68.7%
Dynamic order	73.8%
Proposed framework	79.2%

In the one-incorrect-two-correct setting, the main advantage of the proposed framework is not only correction of the initially incorrect agent but also preservation of the correct agents from being pulled toward the wrong answer in the early stage of the debate.

B. Trajectory of Intermediate Answers by Agent

To examine how other agents' utterances influenced the answers of each agent, we collected the trajectories of intermediate answers for each method. In the legends of the graphs, Alex, Chris, and Jenny denote the three agents in the debate. The agent(s) holding the correct initial answer were Jenny in the two-incorrect-one-correct setting and Chris and Jenny in the one-incorrect-two-correct setting. In the Dynamic order condition, the next speaker was selected based on urgency, whereas in the Fixed order condition speaking rights alternated in the order Alex, Chris, and Jenny.

Fig. 3–5 show the trajectories of agent-wise intermediate-answer accuracy in the two-incorrect-one-correct setting. Under Fixed order (Fig. 3), the accuracy of the initially incorrect

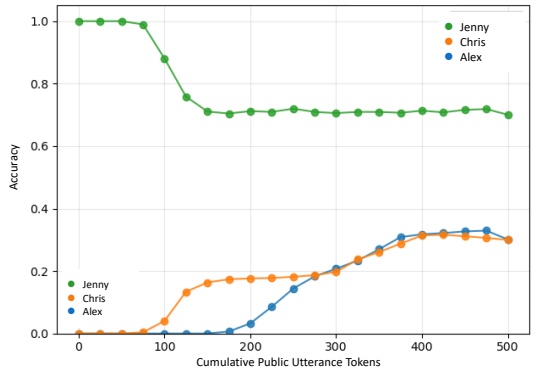

Fig. 3. Trajectory of intermediate-answer accuracy for each agent under Fixed order when two agents were initially incorrect.

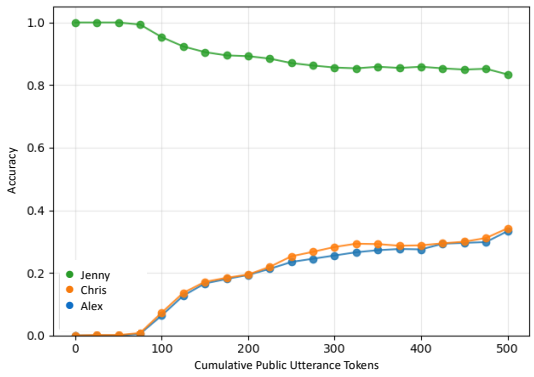

Fig. 4. Trajectory of intermediate-answer accuracy for each agent under Dynamic order when two agents were initially incorrect.

agents Alex and Chris remained close to 0 until the middle stage of the debate (around 200 tokens) and rose only to about 0.3 by the end. In contrast, the initially correct agent Jenny dropped sharply from 1.0 to around 0.7 in the early stage (around 100–150 tokens) and did not recover substantially thereafter. This suggests that the incorrect agents spoke early, formed the shared context, and made the correct agent more vulnerable to the influence of incorrect premises. Under Dynamic order (Fig. 4), Jenny remained around 0.85 while declining only gradually, indicating that the correct side was less likely to be dragged toward the incorrect side than under Fixed order. However, the improvement of Alex and Chris started only after about 100 tokens and plateaued around 0.3, making it difficult for the majority to flip from incorrect to correct. Under the proposed framework (Fig. 5), Jenny stayed at about the same high level as in Dynamic order, while Alex and Chris began improving earlier and reached about 0.4 by the end. The earlier correction of the incorrect side contributed to the higher majority-vote accuracy shown in Table V.

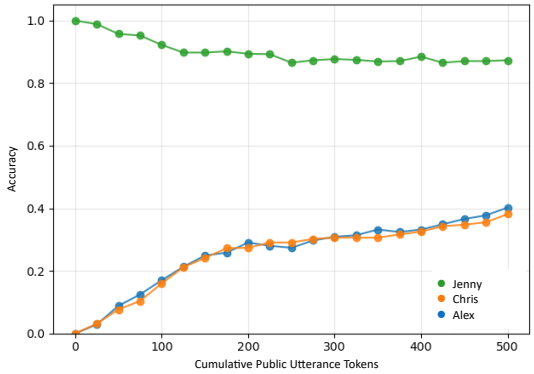

Fig. 5. Trajectory of intermediate-answer accuracy for each agent under the proposed framework when two agents were initially incorrect.

Fig. 6–8 show the corresponding trajectories in the one-incorrect-two-correct setting. Under Fixed order (Fig. 6), the early utterances of the incorrect agent Alex caused the accuracies of the correct agents Chris and Jenny to drop sharply from 1.0 to around 0.6 near 100–200 tokens. Although they later recovered, this early drop lowered the majority-vote accuracy (Table VI). Under Dynamic order (Fig. 7), the decline of Chris and Jenny was more gradual and converged around 0.8, avoiding the sharp deterioration observed under Fixed order. At the same time, Alex also improved from around 100 tokens onward and reached about 0.4 by the end. Under the proposed framework (Fig. 8), Chris and Jenny remained around 0.86–0.9 throughout the debate, while Alex improved step by step from the early stage and also reached about 0.4. In other words, the proposed framework maximized final accuracy by minimizing the correct side's being pulled toward incorrect answers while also accelerating the correction of the incorrect side.

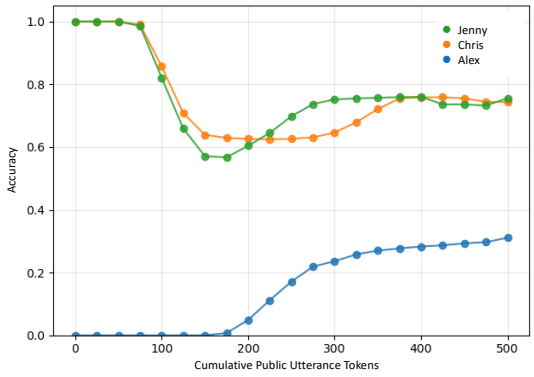

Fig. 6. Trajectory of intermediate-answer accuracy for each agent under Fixed order when one agent was initially incorrect.

VI. Analysis and Discussion

This section analyzes why the proposed framework improved accuracy from the perspective of the debate process, based on the results in Section V.

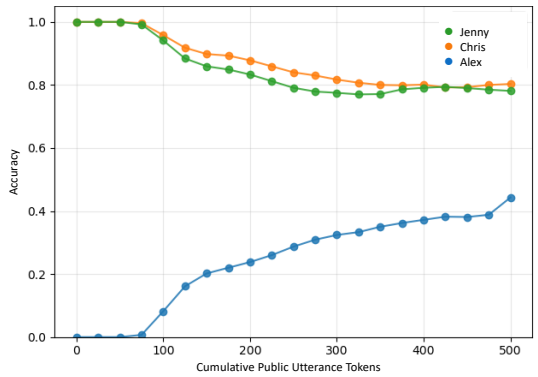

Fig. 7. Trajectory of intermediate-answer accuracy for each agent under Dynamic order when one agent was initially incorrect.

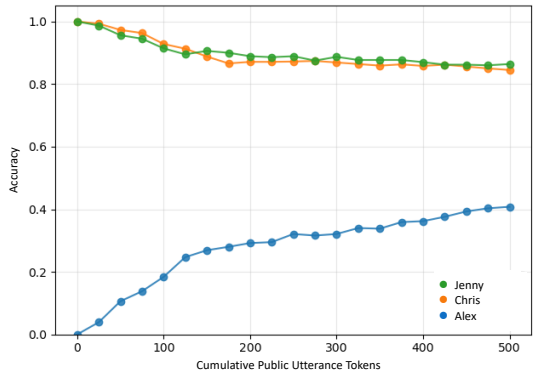

Fig. 8. Trajectory of intermediate-answer accuracy for each agent under the proposed framework when one agent was initially incorrect.

A. Improvement in Final Accuracy

As shown in Table V, in the two-incorrect-one-correct setting the proposed framework achieved a final accuracy of 49.5%, outperforming Fixed order (37.2%) and Dynamic order (43.7%). As shown in Table VI, the proposed framework also achieved 79.2% in the one-incorrect-two-correct setting, outperforming Fixed order (68.7%) and Dynamic order (73.8%).

The two-incorrect-one-correct setting starts with an incorrect majority. When incorrect premises are shared under such a setting, the majority error is likely to be reinforced during the debate. The fact that the proposed method improved accuracy in this setting suggests that it promoted correction on the incorrect side and guided the debate toward flipping the majority from incorrect to correct.

Fig. 1 shows that, under the proposed framework, majority-vote accuracy increased from the early stage of the debate and then remained at a comparatively high level. In contrast, in the one-incorrect-two-correct setting, Fixed order and Dynamic order both exhibited an early drop in accuracy, whereas the proposed framework suppressed that drop (Fig. 2). These results suggest that sentence-by-sentence disclosure and urgency-based interruption allowed intervention and repair to begin before an incorrect claim expanded into a long utterance, thereby suppressing the spread of misinformation while accelerating updates on the incorrect side.

B. Effects of Interruptions and Completion

To further explain why the proposed framework improved accuracy, this subsection compares the effects of 1) interruptions and 2) completion of an utterance on updates to the intermediate answers of the other agents. Here, *completion* refers to the case in which a speaker finishes disclosing the entire utterance (sentence queue) without being interrupted.

1) Evaluation units and scope: An interruption event is defined as the case in which one speaker still has undisclosed sentences remaining in the queue and another agent is selected as the next speaker (Section III). A completion event is defined as the case in which the speaker fully discloses the generated sentence queue and the utterance ends without interruption. In Fixed order and Dynamic order, interruptions do not occur, so every speaking turn is counted as a completion event. In the proposed framework, by contrast, interruptions can occur, so completion events are a subset of the speaking turns that finish without interruption.

For each event, we compared the `Answer` of each agent immediately before the event (just before the speaker change) with the `Answer` after the event (at the time the interrupter finished speaking, in the case of interruption, or when the utterance finished, in the case of completion). The resulting changes in the listeners' `Answer` values were classified into three categories. Because this subsection evaluates the effect of intervention when correct and incorrect answers coexist, cases in which all agents were correct or all agents were incorrect immediately before the event were excluded from the analysis.

The three categories are defined as follows, where s denotes the speaker of the event (either the interrupter or the speaker whose utterance completed) and L denotes the listeners (all agents other than s).

- **Improved**: The speaker s was on the correct side, and at least one listener in L changed from incorrect to correct.
- **Worsened**: The speaker s was on the incorrect side, and at least one listener in L changed from correct to incorrect.
- **Unchanged**: Neither of the above occurred.

For clarity, the percentages reported below are event-level proportions computed over all evaluated events under the stated condition. This event-level view complements final-answer accuracy by showing whether a turn type tends to repair or spread misinformation during the debate process.

2) Effects of interruptions: Under the proposed framework, interruptions yielded an improvement rate of 15.2% and a worsening rate of 7.1% in the two-incorrect-one-correct setting, so improvement exceeded deterioration (Table VII). In the one-incorrect-two-correct setting, the corresponding values were 12.2% and 5.2%, showing the same tendency (Table VIII). Because the worsening rate was lower than the improvement rate in both settings, interruptions can be interpreted as contributing, on average, to suppressing misinformation propagation and advancing repair earlier. The

percentages of improved, worsened, and unchanged reported in this subsection were calculated with all evaluated events as the denominator.

TABLE VII. EFFECTS OF INTERRUPTIONS IN THE PROPOSED FRAMEWORK (TWO INCORRECT, ONE CORRECT)

Metric	Value
Evaluated events	533
Improved	15.2%
Worsened	7.1%
Unchanged	77.7%

TABLE VIII. EFFECTS OF INTERRUPTIONS IN THE PROPOSED FRAMEWORK (ONE INCORRECT, TWO CORRECT)

Metric	Value
Evaluated events	1308
Improved	12.2%
Worsened	5.2%
Unchanged	82.6%

Taken together, Tables VII and VIII indicate that interruption functions more often as an early repair mechanism than as a source of instability under the evaluated setting.

3) Effects of completion: Next, we compared the updates observed immediately after completion events across conditions. In the two-incorrect-one-correct setting, the proposed framework achieved an improvement rate of 14.3% and a worsening rate of 6.5%, again indicating that improvement exceeded deterioration (Table IX). Dynamic order showed a higher improvement rate of 18.1%, but its worsening rate was also high at 11.3%, leaving a smaller net advantage. Under Fixed order, the worsening rate exceeded the improvement rate, resulting in a net negative effect. In the one-incorrect-two-correct setting as well, the proposed framework showed an improvement rate of 10.5% and a worsening rate of 3.7%, indicating a tendency to accumulate improvements while suppressing deterioration (Table X). In the proposed framework, only speaking turns that finished without interruption were counted as completion events.

TABLE IX. COMPARISON OF COMPLETION EFFECTS ACROSS CONDITIONS (COMPLETION WITHOUT INTERRUPTION) IN THE TWO-INCORRECT-ONE-CORRECT SETTING

Condition	Improved	Worsened
Proposed framework	14.3%	6.5%
Dynamic order	18.1%	11.3%
Fixed order	13.6%	18.5%

TABLE X. COMPARISON OF COMPLETION EFFECTS ACROSS CONDITIONS (COMPLETION WITHOUT INTERRUPTION) IN THE ONE-INCORRECT-TWO-CORRECT SETTING

Condition	Improved	Worsened
Proposed framework	10.5%	3.7%
Dynamic order	17.6%	9.2%
Fixed order	13.6%	13.0%

These completion results suggest that the proposed framework does not merely increase the frequency of intervention. Rather, it maintains a more favorable balance between improvement and deterioration even when an utterance is allowed to finish, which is consistent with more stable information sharing than the other two conditions.

4) When interruptions are effective: To analyze at what points in the debate interruptions were effective, we computed the success rates of improvement and deterioration as a function of the cumulative number of public utterance tokens at the time the interruption occurred. Because interruption effects were more clearly visible in the two-incorrect-one-correct setting, we report this analysis for that setting.

Specifically, we binned the cumulative number of public utterance tokens at the time of interruption in increments of 50 tokens and computed the following two conditional probabilities for each bin. Here, *improve_success_rate* is the proportion of interruptions that led to improvement, conditioned on the interrupter being on the correct side, whereas *worsen_success_rate* is the proportion that led to deterioration, conditioned on the interrupter being on the incorrect side.

$$\text{improve\_success\_rate} = \frac{I_{succ}}{I_{succ} + I_{fail}} \tag{1}$$

$$\text{worsen\_success\_rate} = \frac{W_{succ}}{W_{succ} + W_{fail}} \tag{2}$$

Here, I_{succ} denotes the number of cases in which, after an interruption by a correct-side speaker, at least one listener changed from incorrect to correct, and I_{fail} denotes the number of cases in which no such update occurred. Similarly, W_{succ} denotes the number of cases in which, after an interruption by an incorrect-side speaker, at least one listener changed from correct to incorrect, and W_{fail} denotes the number of cases in which no such update occurred.

As shown in Fig. 9, the improve_success_rate was relatively high in the early stage, around 50–199 tokens, and exceeded the worsen_success_rate in many bins. This can be interpreted as indicating that, while the shared context is still short, an interruption can initiate repair immediately after an incorrect premise appears and is therefore more likely to suppress the following spread of misinformation. By contrast, around 300–349 tokens, the worsen_success_rate exceeded the improve_success_rate, suggesting that in the later stage of the debate, interruptions made when few public utterance tokens remain may destabilize consensus formation. This tendency indicates that the effectiveness of interruptions should be controlled according to the stage of the discussion.

C. Limitations

This study has several limitations. First, the evaluation uses a single base model (Llama-3.1-8B-Instruct) and a single benchmark (MMLU), so the results should not be interpreted as evidence of universal effectiveness across models or domains. Second, the experiments fix the number of agents to three and constrain discussion by a 500-token shared public budget; different agent counts or budget regimes may yield different interruption dynamics. Third, urgency is produced by the LLM itself and used as a relative coordination signal, but its calibration is not externally validated in the present study. Fourth, sentence segmentation is performed with SpaCy, and

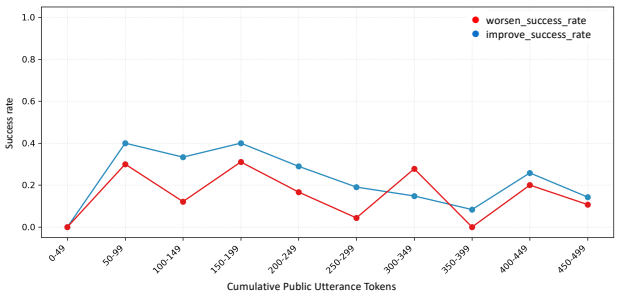

Fig. 9. Improvement and deterioration success rates of interruptions in the proposed framework (50-token bins, two incorrect and one correct).

the robustness of the framework to segmentation errors or alternative segmentation policies is not evaluated. Finally, the initial states are intentionally conditioned to include both correct and incorrect answers so that misinformation propagation can be examined under controlled conditions. This design improves internal comparability across conditions, but it does not fully represent all naturally occurring debate setups.

D. Future Work

Several directions follow from these limitations. First, broader validation is needed across additional LLMs, domain-specific benchmarks, and tasks whose reasoning unfolds differently from multiple-choice question answering. Second, the urgency signal should be calibrated more explicitly, for example by comparing self-reported urgency with externally detected factual conflicts, confidence estimates, or agreement patterns among agents. Third, the framework should be evaluated under larger agent populations and different public-token budgets to clarify how interruption frequency, silence, and consensus stability scale with group size. Fourth, alternative disclosure units and segmentation strategies, including clause-level, discourse-based, or adaptive chunking schemes, should be compared to determine when sentence-level disclosure is most effective. Fifth, because the framework updates action plans at fine granularity, future work should quantify trade-offs among accuracy, generated tokens, number of model calls, and execution time.

In addition, previous studies have reported that assigning personality traits to agents may improve answer accuracy in collaborative tasks and diversify debate [19], [11]. Future work should examine how overall debate accuracy changes when such personality assignments are combined with the interruption- and silence-aware turn-taking control proposed in this study. In particular, it will be important to analyze how differences in personality traits affect the frequency of interruptions and silence, the timing of error correction, and the tendency for unnecessary interruptions to occur. Such analysis may also help realize flexible dialogue in which opinions are presented at effective moments through interruption and silence when the framework is extended to human–LLM discussions.

VII. Conclusion

This study presented an interruptible multi-agent debate framework that combines sentence-level disclosure, urgency-based speaker selection, interruption, and silent turns under a shared public-token budget. The central goal was to reduce the spread of incorrect premises by allowing listeners to intervene before a misleading message expands into a long explanation. Under the evaluated setting, the proposed framework achieved the highest final accuracy among the compared conditions in both the two-incorrect-one-correct and one-incorrect-two-correct settings. Analysis of intermediate answers and event-level updates further suggested that interruption more often supported correction than deterioration. These findings indicate that finer-grained turn control can improve debate stability when correct and incorrect answers coexist. At the same time, the present results are bounded by the evaluated model, benchmark, agent count, and segmentation setting. Broader validation and improved urgency calibration remain important next steps.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Numbers JP23K28377, JP24H00714, JP25K15109, JP25K03190, JP25K03232, JP22K12157 and The Telecommunications Advancement Foundation.

References

[1] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch, "Improving factuality and reasoning in language models through multiagent debate," *arXiv preprint arXiv:2305.14325*, 2023.

[2] S. Bailis, J. Friedhoff, and F. Chen, "Werewolf arena: A case study in LLM evaluation via social deduction," *arXiv preprint arXiv:2407.13943*, 2024.

[3] Y. Feng, Y. Wang, S. Cui, B. Faltings, M. Lee, and J. Zhou, "Unraveling misinformation propagation in LLM reasoning," in *Findings of the Association for Computational Linguistics: EMNLP 2025*. Suzhou, China: Association for Computational Linguistics, Nov. 2025, pp. 11 683–11 707. [Online]. Available: https://aclanthology.org/2025.findings-emnlp.627/

[4] A. Amayuelas, X. Yang, A. Antoniades, W. Hua, L. Pan, and W. Y. Wang, "Multiagent collaboration attack: Investigating adversarial attacks in large language model collaborations via debate," in *Findings of the Association for Computational Linguistics: EMNLP 2024*. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 6929–6948.

[5] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, vol. 50, no. 4, pp. 696–735, 1974.

[6] E. A. Schegloff, G. Jefferson, and H. Sacks, "The preference for self-correction in the organization of repair in conversation," *Language*, vol. 53, no. 2, pp. 361–382, 1977.

[7] G. Li, H. A. A. K. Hammoud, H. Itani, D. Khizbullin, and B. Ghanem, "Camel: Communicative agents for "mind" exploration of large language model society," *arXiv preprint arXiv:2303.17760*, 2023.

[8] Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu *et al.*, "Autogen: Enabling next-gen LLM applications via multi-agent conversation," *arXiv preprint arXiv:2308.08155*, 2023.

[9] C. Qian, W. Liu, H. Liu, N. Chen, Y. Dang, J. Li, C. Yang, W. Chen, Y. Su, X. Cong, J. Xu, D. Li, Z. Liu, and M. Sun, "Chatdev: Communicative agents for software development," *arXiv preprint arXiv:2307.07924*, 2023.

[10] C.-M. Chan, W. Chen, Y. Su, J. Yu, W. Xue, S. Zhang, J. Fu, and Z. Liu, "Chateval: Towards better LLM-based evaluators through multi-agent debate," in *International Conference on Learning Representations*, 2024.

[11] J. Zhang, X. Xu, N. Zhang, R. Liu, B. Hooi, and S. Deng, "Exploring collaboration mechanisms for LLM agents: A social psychology view," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2024, pp. 14 544–14 607.

[12] R. Nonomura and H. Mori, "Who speaks next? multi-party AI discussion leveraging the systematics of turn-taking in murder mystery games," *Frontiers in Artificial Intelligence*, vol. 8, p. 1582287, 2025.

[13] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 24 824–24 837.

[14] Y. Zeng, W. Huang, L. Jiang, T. Liu, X. Jin, C. T. Tiana *et al.*, "S2-mad: Breaking the token barrier to enhance multi-agent debate efficiency," in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Albuquerque, New Mexico: Association for Computational Linguistics, Apr. 2025, pp. 9393–9408.

[15] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, "Measuring massive multitask language understanding," *arXiv preprint arXiv:2009.03300*, 2021.

[16] A. Chowdhery *et al.*, "PaLM: Scaling language modeling with pathways," *arXiv preprint arXiv:2204.02311*, 2022.

[17] OpenAI, "GPT-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[18] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[19] A. Kimura, K. Fukuda, R. Orihara, Y. Tahara, and Y. Sei, "The impact of personality trait integration into LLM agents on collaborative tasks," in *Proceedings of the Annual Conference of the Japanese Society for Artificial Intelligence*, 2025, 3J6-GS-5-04.