



OPEN ACCESS

EDITED BY

Gokhan Tur,
University of Illinois at Urbana-Champaign,
United States

REVIEWED BY

Mert Inan,
Northeastern University, United States
Ishika Agarwal,
University of Illinois at Urbana-Champaign,
United States
Karolina Stanczak,
Montreal Institute for Learning Algorithm
(MILA), Canada

*CORRESPONDENCE

Hiroki Mori
✉ hiroki@speech-lab.org

RECEIVED 24 February 2025

ACCEPTED 16 May 2025

PUBLISHED 17 June 2025

CITATION

Nonomura R and Mori H (2025) Who speaks next? Multi-party AI discussion leveraging the systematics of turn-taking in Murder Mystery games. *Front. Artif. Intell.* 8:1582287. doi: 10.3389/frai.2025.1582287

COPYRIGHT

© 2025 Nonomura and Mori. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](#). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Who speaks next? Multi-party AI discussion leveraging the systematics of turn-taking in Murder Mystery games

Ryota Nonomura and Hiroki Mori*

School of Engineering, Utsunomiya University, Utsunomiya, Japan

Introduction: Multi-agent systems utilizing large language models (LLMs) have shown great promise in achieving natural dialogue. However, smooth dialogue control and autonomous decision making among agents still remain challenging.

Methods: In this study, we focus on conversational norms such as adjacency pairs and turn-taking found in conversation analysis and propose a new framework called “Murder Mystery Agents” that applies these norms to AI agents’ dialogue control. As an evaluation target, we employed the “Murder Mystery” game, a reasoning-type table-top role-playing game that requires complex social reasoning and information manipulation. The proposed framework integrates next speaker selection based on adjacency pairs and a self-selection mechanism that takes agents’ internal states into account to achieve more natural and strategic dialogue.

Results: To verify the effectiveness of this new approach, we analyzed utterances that led to dialogue breakdowns and conducted automatic evaluation using LLMs, as well as human evaluation using evaluation criteria developed for the Murder Mystery game. Experimental results showed that the implementation of the next speaker selection mechanism significantly reduced dialogue breakdowns and improved the ability of agents to share information and perform logical reasoning.

Discussion: The results of this study demonstrate that the systematics of turn-taking in human conversation are also effective in controlling dialogue among AI agents, and provide design guidelines for more advanced multi-agent dialogue systems.

KEYWORDS

turn-taking, conversation analysis, generative AI, LLM-based agent, multi-party conversation

1 Introduction

The emergence of large language models (LLMs) has dramatically enhanced the capabilities of AI agents. With the advent of LLMs such as GPT-3, GPT-4, and LLaMA, we have witnessed the achievement of human-comparable or superior performance across various tasks, including text generation, question-answering, and summarization (Brown et al., 2020; OpenAI, 2023; Touvron et al., 2023; Hugo et al., 2023). The development of AI agents based on these LLMs has gained significant momentum, with promising applications spanning diverse domains such as customer service (Rome et al., 2024), educational support (Jeon and Lee, 2023; Hu B. et al., 2024; Zhang et al., 2024b), gaming environments (Hu S. et al., 2024), economic simulations (Filippas et al., 2024) and creative work assistance (OpenAI, 2022; Anthropic, 2024). Of particular interest is whether AI

agents can exhibit social behaviors similar to those of humans (Lan et al., 2023; Park et al., 2023).

In social interaction, verbal communication plays a central role. Previous studies on the application of LLMs have also revealed that enabling AI agents to chat with each other is an effective approach. For example, *Meta Fundamental AI Research Diplomacy Team (FAIR) et al. (2022)* demonstrated that an AI agent can achieve human-level performance in a complex strategic game by effectively using language for negotiation and coordination. *Qian et al. (2024)* demonstrated that a chat chain between an instructor and an assistant is effective for completing various subtasks in the workflow of software development. *Gu et al. (2024)* proposed a simulation framework for group chats among AI agents, reporting that multifaceted emergent behavior was observed during role-playing scenarios. *Wu et al. (2024)* proposed a platform for LLM applications that supports interaction between LLMs, humans, and tools, where group chats among AI agents are facilitated.

However, text chats are significantly different from human-to-human conversations. It has been claimed that text chat is incoherent, especially due to the lack of interaction management such as simultaneous feedback, which leads to disruption and breakdown of turn-taking and topic management (Herring, 1999). Most AI chat systems employ an even simpler turn-taking model: sending text input from the user initiates the turn transition. This framework does not reflect the properties that human conversation has. For example, chat AIs cannot actively offer topics, initiate conversations, remain silent when other participants are to speak, or withhold from speaking.

As will be discussed in Section 2.2, turn-taking plays a crucial role especially in multi-party conversations. Yet, there have been relatively few studies on such conversation by AI agents. In order to handle multi-party conversations, the problem of selecting the next speaker arises. In the AutoGen platform (Wu et al., 2024), an automatic next-speaker selection mechanism is implemented, where an LLM agent estimates the next speaker's role based on the history of the speaker's role and utterances. However, *Bailis et al. (2024)* pointed out that while this approach is potentially effective, it lacks autonomy for individual agents. Instead, they proposed a dynamic turn-taking system where agents express their desire to speak by bidding.

As *Bailis et al. (2024)* argued, allowing agents to autonomously determine the speaking order could be key to AI agents playing their own social role and having a fruitful conversation. At the same time, however, the order of speaking should not be determined solely by the agents' will. Various perspectives have been proposed on how conversation is structured. For instance, *Clark and Wilkes-Gibbs (1986)* emphasized collaboration in establishing referring expressions, and *Clark and Brennan (1991)* highlighted the importance of grounding for mutual understanding.

While such views frame conversation as a cooperative activity centered on shared knowledge, another influential perspective focuses on the sequential structure of utterances. Sociologists who pioneered conversation analysis schematized such sequential structure as the systematics of turn-taking. Among them, adjacency pairs (Schegloff and Sacks, 1973) are key to understanding the chain of utterances (see Section 2.2 for detailed discussion).

The research question addressed in this study is whether introducing turn-taking systematics such as adjacency pairs, discovered in the research field of conversation analysis, into the next-speaker selection mechanism will have the effect of making LLM-based multi-agent conversations more natural and efficient. *Schegloff (1990)* argued that organization of sequences in turn-taking systematics such as adjacency pairs is the source of coherence in conversation. If so, introducing such a conversational norm into conversations by AI agents is expected to improve the coherence of conversation.

To address this research question, we developed *Murder Mystery Agents (MMAgents)*, a system where multiple AI agents play a deductive tabletop role-playing game called *Murder Mystery*. We chose *Murder Mystery* for several reasons: (1) It demands complex reasoning, information manipulation, cooperation, and negotiation, pushing the boundaries of current NPC (non-player character) capabilities in multi-party conversational games. Simulating NPC players capable of handling such tasks is an open challenge. (2) The game often requires gathering a specific number of human players, which can be difficult; thus, capable NPC substitutes would be valuable. (3) Analyzing the agents' discussions in this structured yet complex environment can provide insights into multi-party argumentation and reasoning processes. Importantly, this study does not aim to directly analyze social behaviors exhibited by LLM-based agents. Rather, it investigates whether the structural mechanisms of dialogue, inspired by findings in conversation analysis—specifically, adjacency pairs—can lead to more coherent and socially plausible interactions among agents. MMAgents incorporates both a self-selection mechanism for autonomous utterances and a next-speaker selection mechanism that detects the first part of adjacency pairs using LLMs to determine the next speaker. Through this design, we aim to explore how introducing conversational norms grounded in turn-taking systematics can enhance the naturalness and efficiency of multi-agent dialogues.

To validate MMAgents, we simulated the discussion phase of “The Ghost Island Murder Case” scenario with four AI agents playing the game's characters. We compared our proposed next-speaker selection mechanism, which leverages adjacency pairs, against two baseline approaches. Experimental results demonstrate the effectiveness of incorporating the principles of conversation found in conversation analysis. Key findings indicate that the proposed next-speaker selection mechanism significantly reduces dialogue breakdowns compared to baseline conditions. Furthermore, evaluations using LLM-as-a-judge and human evaluation tailored to the *Murder Mystery* task show that our proposed framework enhances overall conversational quality, including cooperativeness, diversity, information sharing, and logical reasoning progression within the game context.

2 Background

2.1 LLM-based agents

Numerous LLM-based agents have been proposed so far, as described in Section 1. Of particular interest are multi-agent

systems involving multiple agents. The CAMEL framework (Li et al., 2023) demonstrates how agents with distinct roles can collaborate to solve problems.

Lan et al. (2023) conducted research evaluating social interaction capabilities through multi-agent conversations in the board game Avalon, which requires cooperation and deception among multiple agents. Their study proposed a framework that enables AI agents to make strategic decisions based on previous gameplay experiences, reporting observations of social behaviors such as leadership, persuasion, cooperation, and conflict.

Furthermore, research on AI agents' social behavior, particularly interaction through conversation, continues to evolve. These studies investigate the ability of multiple agents to participate in group chats and discussion scenarios, generating conversations that closely resemble human-to-human interactions (Junprung, 2023; Gu et al., 2024). For example, Park et al. (2023) conducted virtual daily life simulations, analyzing the behavioral patterns of 25 AI agents and their impact on a simulated society. Their study observed information sharing between agents and the formation of novel relationships. There agents have been shown to possess capabilities such as memory, reflection, and planning, enabling more human-like dialogue. This approach contributes to understanding the mechanisms of information exchange and cooperative behavior among agents, potentially offering insights into emergent behaviors in human society.

Research aimed at enhancing LLM-based agents' capabilities is also being actively pursued. For instance, SelfGoal (Yang et al., 2024) proposes automatic generation and updating of sub-goals to achieve high-level objectives. Chain-of-thought prompting (Wei et al., 2022) significantly improves performance on complex reasoning tasks by generating intermediate thought processes. Moreover, ReAct (Yao et al., 2022) proposes an approach alternating between reasoning and action, enhancing agents' ability to solve problems incrementally while interacting with their environment.

Overall, most existing research deals with one-to-one interactions or simplified turn-taking mechanisms, failing to address the natural flow of conversation that occurs in groups of three or more participants.

2.2 Turn-taking

In human conversation, there exists a fundamental constraint where typically only one person speaks at a time. This constraint stems from the physical limitations of speech communication, as simultaneous speech by multiple participants leads to interference, making comprehension difficult. For efficient communication, speakers must smoothly alternate turns while minimizing silent intervals between utterances. To meet this requirement, humans have naturally developed turn-taking systems through social interaction.

Turn-taking, where dialogue participants take turns to speak, forms the foundation of smooth communication. Through analysis of spontaneous conversation recordings, conversation analysts like Sacks, Schegloff, and Jefferson systematically described this

phenomenon and identified the following rules (Sacks et al., 1974):

1. If the current speaker designates the next speaker by using a "current speaker selects next" technique (e.g., at the first pair part of an adjacency pair Schegloff and Sacks, 1973), the selected participant has both the right and obligation to become the next speaker. (Current Speaker Selects Next)
2. If the current speaker does not designate the next speaker, other participants can spontaneously initiate speech. (Self-Selection)
3. If no one begins speaking, the current speaker can continue.

Unlike dyadic conversations where speaker and listener roles are clearly defined, multi-party conversations involve multiple participants, necessitating the use of gaze direction and verbal addressing to designate the next speaker (Sacks et al., 1974). Adjacency pairs, the basic units of conversation, consist of paired utterances such as [question-answer] and [invitation-acceptance/rejection]. The initial utterance is referred to as the first pair part, and the responding utterance as the second pair part. First pair parts like "I'd like to purchase this item (request)" generate an obligation for a specific type of second pair part [in this case, "Certainly (acceptance)" or "We're sold out (rejection)"]. An inappropriate second pair part or lack of response suggests either a communication error or implies a reason for the inability to respond. Such functional binding is called conditional relevance (Schegloff, 1968). When the current speaker addresses a question to another one, the addressee is not only obligated to take the turn, but also to speak something relevant to the question. In multi-party conversations, the first pair part of adjacency pairs often involves this "current speaker selects next" technique (Sacks et al., 1974).

Humans dynamically create conversations as collaborative acts among participants using this turn-taking system. In contrast, current AI agents struggle to autonomously engage in such flexible and immediate interactions. Therefore, implementing turn-taking mechanisms in AI agents may enable more natural and smooth dialogue.

Research on turn-taking in multi-agent systems remains extremely limited. As noted in Section 1, recent work by Bailis et al. (2024) represents one of the few efforts to explicitly address this issue. While several studies have explored multi-party conversational agents, they tend to overlook the central question of how agents should coordinate speaker transitions—a crucial component for maintaining coherent and socially plausible interactions. To date, the relevance and technical challenges of turn-taking in multi-agent dialogue have not been sufficiently recognized or systematically investigated in the literature.

2.3 Murder Mystery

Murder Mystery is a reasoning-type table-top role-playing game in which players play the roles of characters within a story, aiming to either identify the murderer or, if playing as the murderer, to avoid detection. The game's progression heavily relies on players sharing information through conversation, including evidence gathered from crime scene investigations and character-specific knowledge. Furthermore, Murder Mystery

assigns different missions to each player. Players may need to cooperate or deceive others to accomplish these missions. This requires not merely intelligence but also human-like social behaviors such as teamwork, persuasion, negotiation, and deception. Successfully replicating these behaviors in AI agents could lead to significant advances in artificial intelligence research.

There has been one attempt to make AI agents play Murder Mystery games (Junprung, 2023). In this prior research, a detective agent poses the same questions to five agents, including the murderer. After all five responses are collected, the detective agent responds and asks another question. This process is repeated N times, after which the detective agent attempts to identify the murderer. This approach is termed “one-to-many simulation.” While the simulation successfully identifies the murderer, this method does not accurately reflect real Murder Mystery gameplay, where all players except the murderer must develop their own theories to identify the murderer. While this approach is referred to as “many-to-many,” it could not be implemented due to OpenAI’s input token limitations. Therefore, this research aims to develop an agent framework capable of either reasoning or concealing information about the murder through autonomous conversation, similar to human players.

3 Conversational agents simulating human multi-party conversation

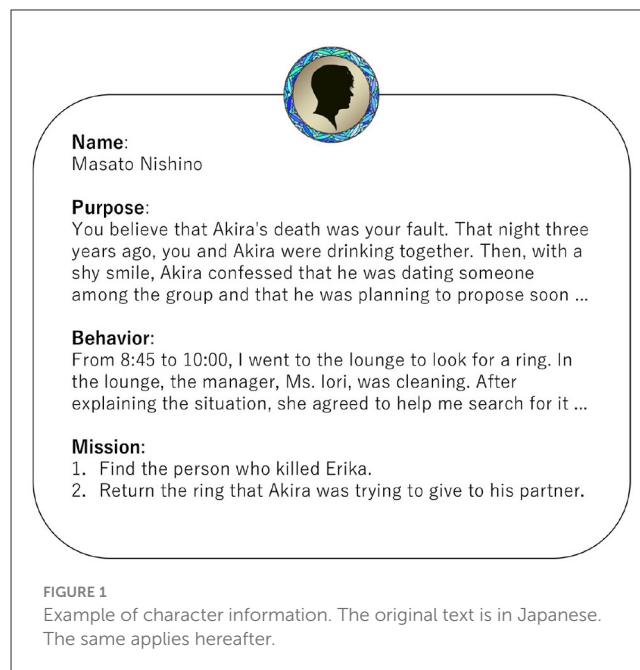
Building upon the characteristics of Murder Mystery games discussed in Section 2.3, this section details the design philosophy and technical components of MMAgents (Murder Mystery Agents), a system developed to facilitate autonomous game progression. MMAgents is designed to simulate multi-party human conversations, enabling multiple AI agents to not only cooperate but also engage in complex conversations involving competition and bargaining to advance the Murder Mystery game.

3.1 Component

3.1.1 Character setting

In Murder Mystery games, before the game begins, the game master provides players with character sheets. Each character sheet contains information necessary for players to portray their characters, including background, personality, objectives, and actions on the day of the incident. Players read and understand this information and play the character to talk and explore.

The approach of having LLMs roleplay characters and evaluating their performance has been reported in several studies (Shanahan et al., 2023; Shao et al., 2023; Wang et al., 2024; Lu et al., 2024). As shown in Figure 1, MMAgents structures each agent’s prompt beginning with the character’s name, followed by descriptions of their objectives, actions, and missions to accomplish. For example, the character Masato Nishino’s information includes crucial background details such as memories of his close friend Akira who passed away three years ago, and romantic feelings expressed that night. The information also includes specific incident-related actions, such as his behavior in



the lounge the previous day and conversations with the inn’s manager. Furthermore, character-specific missions are established, such as “finding Erika’s murderer” and “returning the ring that Akira intended to give to his lover.”

In this way, each agent is provided with character information containing distinct backgrounds and objectives, which guides their decision-making and dialogue. Only surface-level information about other characters is shared, and this information asymmetry implements the elements of information gathering and strategic interaction inherent in Murder Mystery games.

3.1.2 Memory

For LLM-based agents, memory management mechanisms are crucial components for generating more natural and consistent responses in user interactions (Zhang et al., 2024a; Zhong et al., 2024; Modarressi et al., 2023). This is equally important in agent-to-agent dialogue (Park et al., 2023). To create systems like Murder Mystery, where multiple agents engage in complex discussions over extended periods, it is essential to appropriately store past statements and acquired information, and recall them at necessary moments.

Building upon extensive research on memory modeling in cognitive science which has established foundational frameworks for distinguishing between short-term and long-term memory and understanding their interaction (Atkinson and Shiffrin, 1968; Baddeley and Hitch, 1974; Miller, 1956; Tulving, 1972), we designed our Memory module by adapting the practical architecture proposed by Park et al. (2023). This research manages agents’ memory across three distinct layers. First, there is a memory named *History* that is shared by all agents, which maintains the past k turns of conversation as shown in Equation 1. History is used to maintain conversational context and track recent dialogue flow.

$$\text{history} = \{u_{n-k+1}, u_{n-k+2}, \dots, u_n\}, \quad (1)$$

where u_i represents the i -th utterance.

Second, each agent maintains a short-term memory, named *shortTermHistory*. This consists of a history of thoughts generated by the *think()* function detailed in Section 3.2.1, and maintains agent-specific policies and intentions, as shown in Equation 2,

$$\text{shortTermHistory} = \{t_{n-k+1}, t_{n-k+2}, \dots, t_n\}, \quad (2)$$

where t_i represents the i -th thought. The *shortTermHistory* enables agents to maintain consistency in their reasoning and intentions.

Furthermore, each agent maintains a long-term memory, named *longTermHistory*, in which utterance content is normalized using LLMs, and important knowledge and information is extracted and stored in a database, as formulated in Equation 3. Supplementary Figure S1 demonstrates the process of information extraction and normalization in *longTermHistory*. This example illustrates the process of extracting important information from unstructured speech text by Kozue Taniguchi and storing it as structured knowledge. This normalization process facilitates later retrieval and reference by extracting important facts and information from unstructured text in a bullet-point format.

$$\text{longTermMemory} = \{k_1, k_2, \dots\} \quad (3)$$

When generating new utterances, the previous utterance u_{t-1} is converted into an embedding vector $E(u_{t-1})$, and the cosine similarity shown in Equation 4 is calculated with each vector $E(k_i)$ of the embedded knowledge stored in *longTermHistory* to retrieve relevant past memories.

$$\cos(E(u_{t-1}), E(k_i)) = \frac{E(u_{t-1}) \cdot E(k_i)}{|E(u_{t-1})| |E(k_i)|} \quad (4)$$

The calculated similarities are sorted in descending order, and normalized knowledge (k_i) corresponding to the top l vectors is selected. This enables efficient recall of past memories relevant to the current context, which agents can utilize for reasoning and utterance generation. These three layers of memory systems each have different time scales and purposes. History maintains the flow of recent conversations, *shortTermHistory* retains each agent's thought processes, and *longTermHistory* stores important facts and information. By incorporating these memories into prompts, agents can generate contextually appropriate utterances and maintain consistent conversations.

3.2 Turn-taking system

The turn-taking system is potentially a crucial element for achieving natural dialogue among multiple agents. In conventional multi-agent dialogue systems, speaking turn was often predetermined or randomly assigned. In this research, based on Sacks et al. (1974)'s conversation analysis theory discussed in Section 2.2, we implemented two characteristic turn-taking mechanisms from natural human conversation in MMAGents: "Self-Selection" and "Current Speaker Selects Next". This enables natural turn-taking that reflects the agents' personalities and intentions. The pseudocode for this algorithm is shown in Figure 2. This subsection details the important modules of the turn-taking algorithm.

3.2.1 think()

At the beginning of each turn, agents execute an action called *think()*. Based on the provided character data, *think()* generates thought, which represents the plan for the next utterance or action aimed at achieving their mission. Simultaneously, it decides whether to take the action of "speak" or "listen". This selection is implemented with the assumption that it is determined by considering other agents' utterances and the urgency of their own thought content. Furthermore, it outputs an importance as an integer from 0 to 9. This scale is adopted following Park et al. (2023), and is designed to reproduce the Self-Selection mechanism in conversation and is presumed to be determined based on factors such as relevance to the mission, consistency with current conversational context, urgency of the utterance content, and character personality. The prompt used for *think()* is shown in Supplementary Figure S2.

Figure 3 shows an example where four agents execute *think()*. In this example, Kozue Taniguchi and Yukiko Shiraishi chose "speak", with Kozue Taniguchi in particular outputting a high importance value. This suggests that Kozue Taniguchi judged her utterance to be significant for the conversation's development.

3.2.2 selectMostImportant()

The *selectMostImportant(agents)* is a speaker selection algorithm that implements the Self-Selection mechanism. This algorithm processes differently based on the number of agents who have selected "speak." When only one agent selects "speak", that agent naturally becomes the speaker. This is the simplest case of Self-Selection. Conversely, when multiple agents select "speak", their importance values are compared, and the agent with the highest value becomes the speaker. This represents the turn-taking systematics of "the first person to start speaking becomes the speaker", expressed numerically through importance values. In cases of tied importance values, random selection is used to represent the uncertainty of turn-taking in actual conversations. Furthermore, when all agents select "listen," the previous speaker continues speaking. This implements the turn-taking systematics that "when the current speaker does not select the next speaker, they retain the right to continue speaking". However, in the first turn at the start of the dialogue, the speaker is determined randomly.

In the example shown in Figure 3, although both Kozue Taniguchi and Yukiko Shiraishi selected "speak", Kozue Taniguchi is chosen as the next speaker due to her higher importance value.

3.2.3 speak()

The selected agent as speaker generates an utterance using the prompt shown in Figure 4. This prompt consists of the character data shown in Figure 1 and the three types of memory (History, *shortTermHistory*, *longTermHistory*) explained in Section 3.1.2. This enables natural utterances that consider the agent's personality, past conversation content, and policies.

Algorithm 1: Turn-taking system

```

nextSpeaker ← nil
while true do
  currentSpeaker ← nextSpeaker
  for agent in agents do
    thought, action, importance ← agent.think()
    agent.thought ← thought
    agent.action ← action
    agent.importance ← importance
  if currentSpeaker is nil then
    currentSpeaker ← selectMostImportant(agents)
  for agent in agents do
    if agent is currentSpeaker then
      utterance ← agent.speak()
      agent.shortTermHistory.append(utterance)
    else
      agent.shortTermHistory.append(agent.thought)
  history.append(utterance)
  longTermHistory.append(knowledgeNormalization(utterance))
  nextSpeaker ← detectDesignation(utterance)

```

FIGURE 2
Algorithm for turn-taking system.

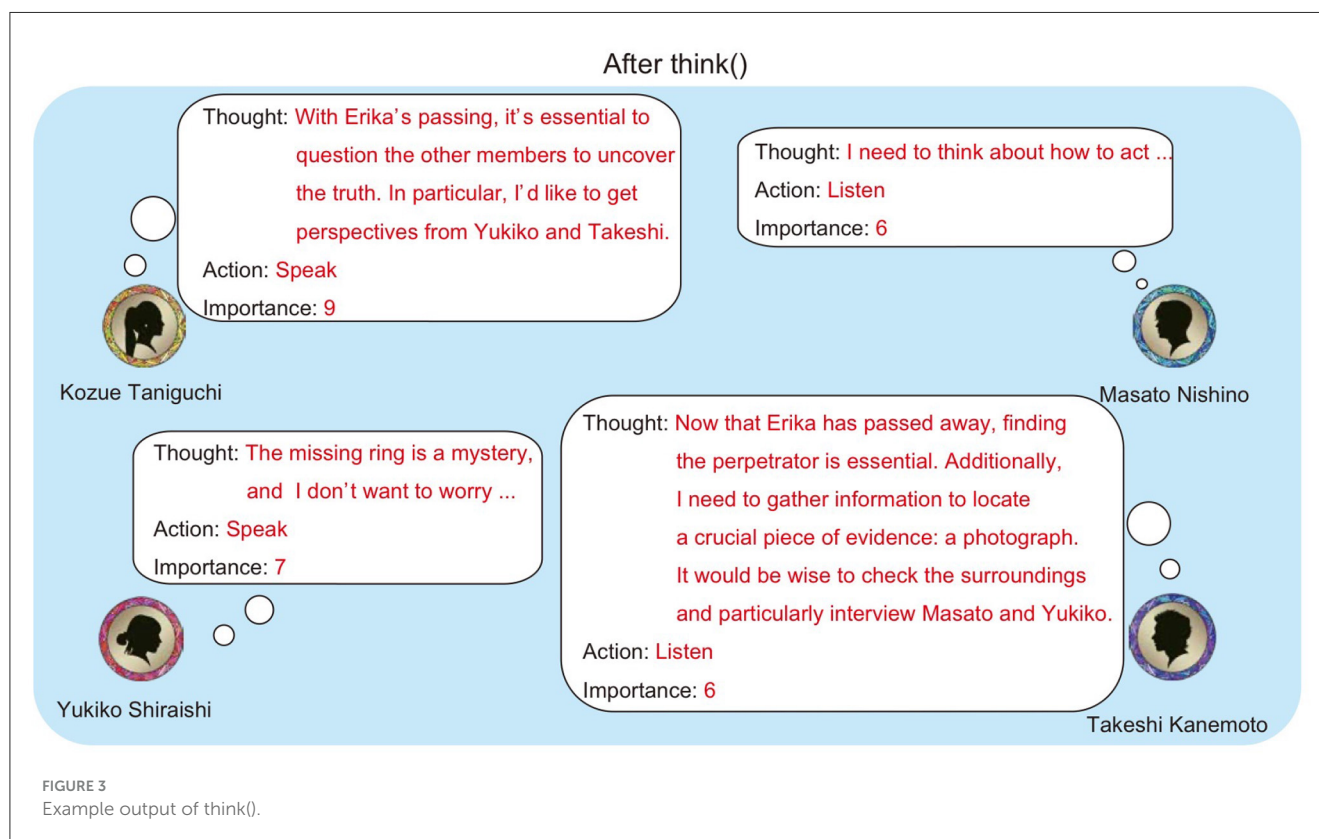
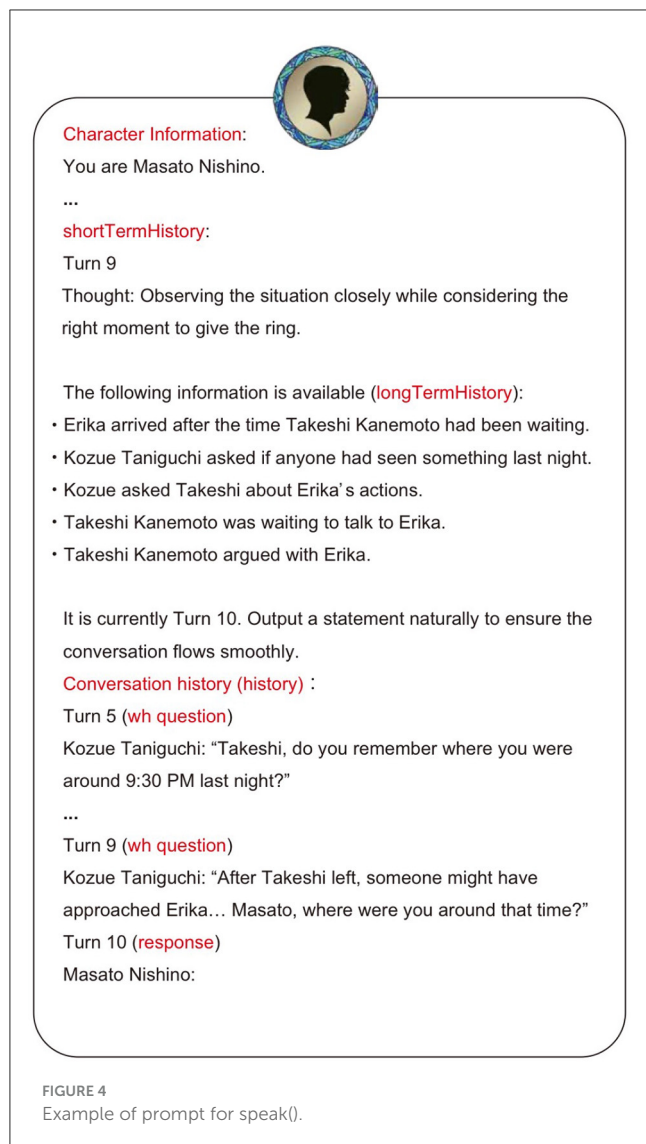


FIGURE 3
Example output of think().



3.2.4 detectDesignation()

detectDesignation() is a mechanism that detects whether the current speaker has explicitly designated the next speaker. This process uses the LLM to determine if a first pair part of an adjacency pair is present in the previous turn's utterance. When a first pair part is detected, it simultaneously classifies its type (Yes/No question, addressing, etc.) and estimates the agent addressed by the utterance. In the example shown in Figure 5, Kozue Taniguchi asks Masato Nishino "Where were you at that time?". When this utterance is input to detectDesignation(), the LLM outputs the detected type of first pair part (wh question) and the predicted next speaker (Masato Nishino).

Then, by incorporating the type of the corresponding second pair part into the prompt used in the following speak(), the agent designated as the next speaker is obligated to respond to the previous turn's utterance. For example, in Turn 10 of the conversation history in Figure 4, a constraint of "(response)" is imposed on the next speaker's utterance, because the previous utterance was a first pair part (wh question). This achieves

coherency in adjacent utterances while maintaining natural conversation flow.

4 Experiments and evaluations

4.1 Experiments

To validate the effectiveness of the proposed MMAgents, we conducted conversational simulations using a commercially available Murder Mystery scenario titled "The Ghost Island Murder Case".¹ This scenario was selected because it features characters with well-defined roles and positions, while maintaining a moderate difficulty level for non-murderer characters, with logical deductions that are challenging yet solvable. "The Ghost Island Murder Case" begins with a story of former college tennis team members reuniting on an isolated island after three years. The scenario features the following four characters:

- Kozue Taniguchi (female): A boyish character with a straightforward personality.
- Masato Nishino (male): An energetic character. Endearing, but sometimes fails to read the room.
- Yukiko Shiraishi (female): A caring, big-sister type character in the group, though she has a tendency to overthink.
- Takeshi Kanemoto (male): A sincere character despite his flashy appearance.

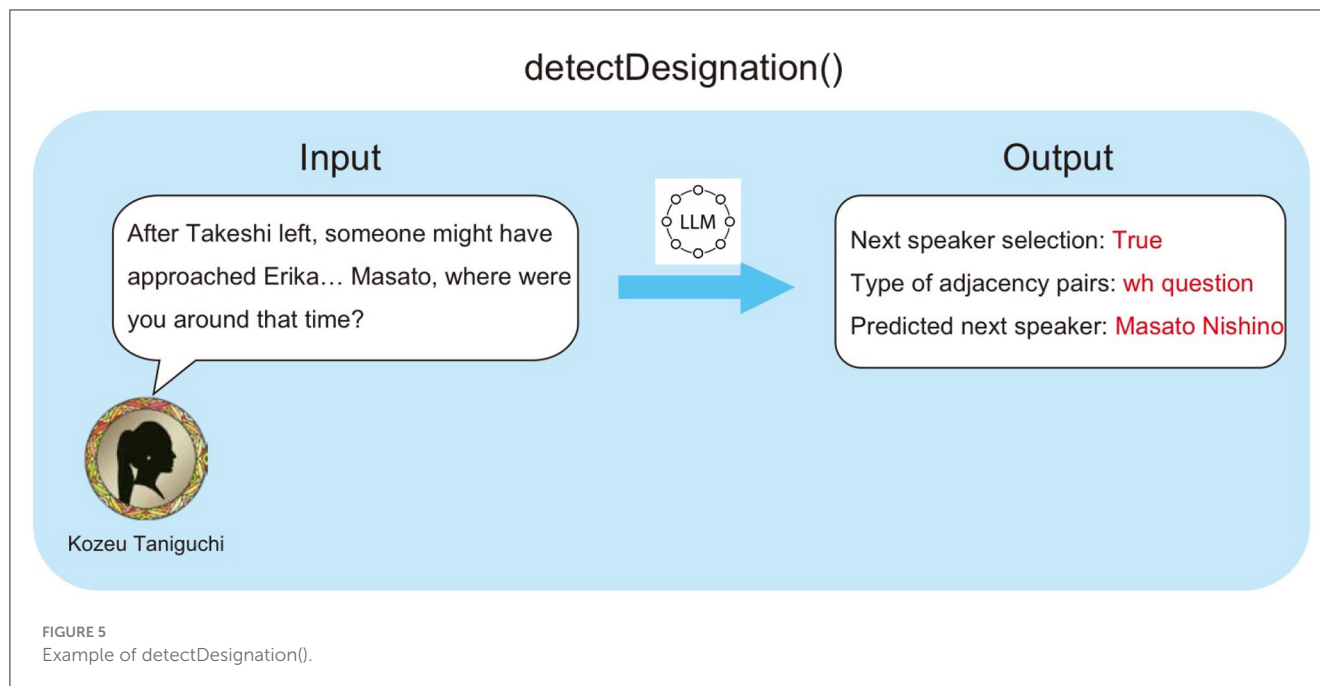
While the scenario consists of multiple phases (exploration phase for information gathering, private conversation phase, discussion phase, reasoning phase, etc.), our experiment focused solely on the discussion phase. This choice was primarily motivated by our aim to evaluate the effectiveness of MMAgents' core functionality: human-like turn-taking. We determined that the discussion phase, with its active dialogue and exchange of opinions between participants, would be optimal for assessing the performance of our proposed method.

In our experiments, we employed multiple large language models. GPT-4o was utilized for detectDesignation() and speak(), as these tasks require sophisticated context understanding and natural speech generation. Conversely, GPT-3.5-turbo was employed for simpler tasks such as knowledge normalization (longTermHistory) and think() to optimize computational costs. To accommodate the input token limitations of LLMs, we set the retained turns for History and shortTermHistory to five turns, while longTermHistory was configured to select the top five entries based on similarity scores.

To evaluate the proposed method, we conducted experiments under the following three conditions:

- EQUAL: The participants have equal opportunity to speak.
- SS: The next speaker always Selects Self.
- CSSN-or-SS: Current Speaker Selects Next, otherwise the next speaker Selects Self.

¹ <https://booth.pm/ja/items/1624107>



In the EQUAL condition, the order of speaking is randomly determined each round. This ensures that the number of each participant's utterances is equal, while avoiding potential order effects. In the CSSN-or-SS condition, the turn-taking system described in Section 3.2 determines the speaking order. The SS condition is the same as the CSSN-or-SS condition except that it does not have the detectDesignation() mechanism, which is used for speaker selection in the next turn.

For each condition, we generated 50 sets of 10-turn conversations. The results were then evaluated using the evaluation methods described in the following subsection, enabling a statistical analysis of the effectiveness of our proposed approach.

4.2 Evaluations

To evaluate the conversations generated by MMAgents, we adopted the following three approaches:

1. Analysis of dialogue breakdown: To assess the naturalness of generated conversations, we employed LLMs to analyze and evaluate the number of utterances that led to dialogue breakdowns (Higashinaka et al., 2022).
2. LLM-as-a-Judge: We defined three metrics—coherence, cooperation, and conversational diversity—and evaluated them using score-based LLM as the judging methodology (Li et al., 2024; Kocmi and Federmann, 2023).
3. Human evaluation: We established original evaluation criteria focusing on Murder Mystery game progression and information sharing between agents. These criteria comprehensively assess the agents' reasoning capabilities and information-gathering abilities through analysis of conversations generated by MMAgents.

These methods were selected for their respective strengths. LLM-based evaluations (1 and 2) enabled consistent analysis across a large number of dialogues, while human evaluation (3) was necessary to assess aspects that exceed the capacity of LLMs, such as complex reasoning and long-context interpretation required in the Murder Mystery. It should be noted that this study does not aim to investigate the correlation or interchangeability between human and LLM-based evaluations, but rather to use them complementarily to assess different dimensions of dialogue quality.

We employed GPT-4 for both the dialogue breakdown analysis and the LLM-as-a-judge, as it is empirically known to exhibit strong judgment capabilities. In addition, to examine inter-rater agreement among various LLMs, we conducted the same evaluations using Claude and Gemini.

4.2.1 Analysis of dialogue breakdown

Evaluating conversational naturalness is crucial, but difficult to achieve. The evaluation of naturalness is inherently subjective, heavily dependent on evaluators' perspectives and prior experiences. Even when different evaluators assess the same conversation, their evaluations may not align, making it difficult to establish standardized evaluation criteria. Therefore, rather than directly evaluating conversational naturalness, our research adopts an indirect approach by evaluating the degree of dialogue breakdown. Specifically, we employ the "classification of utterances that lead to dialogue breakdowns" proposed in dialogue systems research (Higashinaka et al., 2022). Among these types, we use LLMs to analyze items corresponding to response and context-level errors shown in Supplementary Table S1. For the analysis, we input 10-turn conversation samples generated by MMAgents into the LLM, which then identifies utterances corresponding to the categories in Supplementary Table S1 as breakdown utterances (B) and others as non-breakdown utterances (NB). This

system prompt: You are an expert in conversation analysis. Analyze whether the conversation is breaking down according to the specified conditions, and respond in the following JSON format.

prompt:

Conditions for dialogue breakdown:

- **Ignore question:** The question is ignored.
- **Ignore request:** The response does not appropriately address a request like "Please do xxx."
- **Ignore suggestion:** The response does not appropriately address a suggestion like "Let's do xxx."
- **Ignore greeting:** Failing to properly respond to a greeting.
- **Ignore expectation:** The response addresses the previous statement, but does not contain the expected content.
- **Unclear intention of utterance:** The relationship between the response and the topic is understood, but the intention behind the statement is not clear and cannot be inferred from the context.
- **Topic-change error:** Shifting to a different topic without explanation from the topic the other party was discussing. However, this does not apply if the adjacent utterance does not have a clear topic or if the previous utterance marks the end of a topic.
- **Repetition:** Simply repeating something already stated or persistently asking the same question.
- **Lack of information:** The statement lacks necessary elements like modifiers, subject, or object, making it difficult to understand in context.
- **Self-contradiction:** The statement contradicts something the speaker has previously said.
- **Interlocutor contradiction:** The statement contradicts what the other party has said or includes questions that negate the content of their statement.

Detect and output all conditions for dialogue breakdown included in the input. If none of the conditions apply, answer "none".

FIGURE 6
Prompt for the analysis of dialogue breakdown.

process is repeated 50 times, and then conversational naturalness is quantitatively evaluated through statistical analysis of the distribution of utterances identified as B. We utilized GPT-4 for this analysis, with the prompt shown in Figure 6.

4.2.2 LLM-as-a-Judge

A new approach called "LLM-as-a-Judge" has emerged for evaluating natural language processing tasks (Li et al., 2024; Kocmi and Federmann, 2023; Gao et al., 2023; Chiang and Lee, 2023). This rapidly evolving methodology is increasingly being recognized as an alternative to traditional human evaluator-dependent methods. The fundamental concept of the "LLM-as-a-Judge" approach involves inputting some text or conversation to be evaluated into LLMs and having them perform evaluations based on specific criteria or metrics. The primary advantage of this method lies in its ability to analyze large volumes of data efficiently and consistently without requiring human evaluators.

We employ LLMs to evaluate the quality of generated conversations using three metrics: coherence, cooperativeness, and diversity. Coherence evaluates the logical flow and absence

of contradictions in conversations, with scores ranging from 1 (contradictory and illogical) to 5 (consistent and logical). Cooperativeness evaluates how collaboratively participants engage in information exchange and solving problems, with scores ranging from 1 (uncooperative) to 5 (cooperative). Conversational diversity evaluates the absence of repetitive content and the presence of varied opinions and perspectives, with scores ranging from 1 (no diversity) to 5 (high diversity). Coherence indicates the logical flow of conversation, Cooperativeness reflects the quality of participant interactions, and diversity represents the richness and depth of the conversation. In the evaluation process, each conversation sample is input into the LLM, which outputs scores from 1 to 5 for each of the three metrics mentioned above. We utilized GPT-4 for this evaluation.

4.2.3 Human evaluation

To evaluate the quality and effectiveness of conversations in the Murder Mystery scenarios, the authors developed original evaluation criteria and conducted detailed evaluations of each conversation from the perspectives of information-sharing efficiency and discussion progression. Our evaluation criteria were designed based on the hypothesis that smooth conversation facilitates logical discussion, ultimately leading to the game's objective of solving the case. A portion of these evaluation criteria is shown in Figure 7.

The evaluation of information-sharing efficiency measures the activity of information exchange, which forms the foundation for in-depth discussion. Specifically, points are awarded when character-specific information is appropriately disclosed during conversation. This quantitatively evaluates the quality of information sharing that serves as the basis for case-solving reasoning.

The evaluation of discussion progression measures the development of reasoning based on shared information and the progress toward solving the case. Points are awarded when characters demonstrate logical reasoning and insights, or when significant facts are revealed. This enables quantitative evaluation of progress toward the task of uncovering the truth behind the case.

This methodology enables systematic evaluation of the entire process, from information sharing through logical reasoning to case resolution. In particular, by considering the specific characteristics of murder mysteries, we can more concretely verify the effectiveness of our proposed method.

5 Results

We compared three types of generated conversations: those with equal speaking turns and opportunities (EQUAL) as detailed in Section 4.1, those generated using only the Self-Selection mechanism (SS), and those generated using our proposed approach incorporating the Current Speaker Selects Next mechanism (CSSN-or-SS). Examples of generated conversations are shown in Figures 8–10.

Examining the EQUAL condition example in Figure 8, in Turn 1, Masato asks Takeshi "Did something happen?" In Turn 2, Kozue follows with another question to Takeshi. Turn 3 similarly

Guidelines for human evaluation

Scoring criteria (Points can be awarded for any speaker's statement)

- **1 Point:** Information directly given in the prompt is mentioned.
- **3 Points:** Information is directly given but touches on sensitive details that the character wants to hide or is central to solving the case or mission.
- **5 Points:** Achieving a mission unrelated to the case.
- **10 Points:** The truth of the case is revealed.

1. The murderer of Erika: Taniguchi Kozue (10 points)
 2. The murderer of Akira: Takeshi (10 points)
 3. The person argued with is Takeshi (3 points)
 4. Erika is an accomplice (3 points)
 5. About the evidence photo (3 points)
 6. The exchange of the ring (5 points)
 7. Notices the ring swap (3 points)
 8. The truth about the lipstick (Yukiko wrote with the lipstick) (3 points)
- Information included in the prompt (1 point for each)
- Kozue Taniguchi
- 20:30
9. I went to the lounge to get a drink because I was thirsty.
- ...

FIGURE 7
Guidelines for human evaluation.

contains a question directed at Takeshi. Finally, in Turn 4, Takeshi responds to these questions. This example clearly demonstrates the phenomenon of conversational inefficiency caused by fixed speaking order. Specifically, when an utterance that corresponds to the first pair part of an adjacency pair occurs, the generation of a second pair part requires waiting until one's turn comes around. This problem increases the cognitive load for participants tracking conversation context and disrupts the natural flow of dialogue, resulting in inefficient conversation.

In the SS condition example shown in Figure 9, Turn 1 shows Takeshi asking all participants "Where was everyone?" In Turn 2, Masato explains his behaviors in response to the Turn 1 question. In Turn 3, Kozue asks Yukiko "What were you doing?" However, from Turn 4 to Turn 7, Kozue continues to ask questions. This example demonstrates the phenomenon of turn monopolization caused by importance-based turn allocation. When an agent with high importance scores (in this case, Kozue) dominates the conversation, it limits other participants' opportunities to speak and prevents proper generation of second pair parts of adjacency pairs. This problem disrupts conversational balance and undermines the fairness of information exchange and the bidirectional nature of dialogue.

In the CSSN-or-SS condition example shown in Figure 10, Turn 1 shows Kozue asking all participants "Did anyone see anything

in the lounge last night?" In Turn 2, Masato responds to Turn 1's question "I was in the lounge but didn't see anything." In Turn 3, Kozue uses a sequence-closing third (Schegloff, 2007) saying "I see, thank you" to conclude the conversational sequence with Masato. Kozue then addresses Yukiko, asking a similar question based on memory that Yukiko had said she would "go to the lounge." In Turn 4, Yukiko provides the second pair part of the adjacency pair, responding "I didn't meet anyone at that time." In subsequent turns, the conversation flow continues with Kozue addressing specific participants with questions and appropriate agents providing responses.

In our analysis of 500 think() calls across experiments, no instance occurred where all agents simultaneously selected "listen." This absence is attributed to the game's design: each agent is assigned missions achievable only through active dialogue (e.g., information sharing or deception). Consequently, agents are inherently motivated to prioritize "speak" actions to fulfill objectives, preventing conversational stalls.

Our analysis revealed a skew in importance scores toward higher values. The distribution per agent was as follows:

- Kozue Taniguchi: 7 (2%), 8 (48%), 9 (50%)
- Masato Nishino: 7 (16%), 8 (73%), 9 (12%)
- Yukiko Shiraishi: 7 (8%), 8 (69%), 9 (23%)
- Takeshi Kanemoto: 7 (6%), 8 (53%), 9 (41%)

These patterns reflect character backgrounds. Kozue and Takeshi possess critical secrets directly related to the murder case, which likely led to higher urgency scores as they were motivated to take initiative in conversation to conceal those secrets. Yukiko, who holds a relatively minor secret not central to the crime, exhibited moderately high urgency. In contrast, Masato, who has little information directly relevant to the case, showed comparatively lower urgency.

Figure 11 shows the analysis results of dialogue breakdown described in Section 4.2.1. In both the EQUAL and SS conditions, the number of utterances that led to dialogue breakdown per 10 turns showed a wide distribution from one to eight utterances. Conversely, the CSSN-or-SS condition showed a narrow distribution centered around one utterance. A Kruskal-Wallis test revealed significant differences between conditions ($\chi^2 = 42.171$, $p < 0.001$). Dunn's multiple comparison test (with Bonferroni correction) showed that the CSSN-or-SS condition significantly reduced utterances that led to dialogue breakdowns compared to the EQUAL condition ($p < 0.001$) and the SS condition ($p < 0.001$).

Figure 12 shows the LLM-as-a-Judge evaluation results described in Section 4.2.2. For the metrics of coherence, cooperativeness, and diversity, the EQUAL condition showed peaks at score 4, while the SS condition showed wide distributions from scores 2 to 4. The CSSN-or-SS condition distributed across scores 4 and 5, with diversity showing a notable peak at score 4. A Kruskal-Wallis test revealed significant differences between conditions for all metrics (coherence: $\chi^2 = 51.784$, $p < 0.001$; cooperativeness: $\chi^2 = 56.718$, $p < 0.001$; diversity: $\chi^2 = 52.973$, $p < 0.001$). Dunn's multiple comparison test (with Bonferroni correction) showed no significant difference between CSSN-or-SS

-
- 1 Masato: "I saw Takeshi running from the beach at 22:05 last night, did something happen?"
 - 2 Kozue: "Takeshi, were you the one talking with Erika around 21:45 yesterday?"
 - 3 Yukiko: "Takeshi, where did you go after talking with Erika?"
 - 4 Takeshi: "After talking with Erika, I returned to the cottage. Yukiko, I remember meeting you. I wasn't feeling well, so I went back to my room right after."
 - 5 Kozue: "Takeshi, you felt unwell after talking with Erika, did something happen?"
 - 6 Takeshi: "Erika brought up Akira and said various things I didn't want to remember. It made me feel a bit confused."
 - 7 Masato: "Takeshi, why were you running from the beach?"
 - 8 Yukiko: "Takeshi, did you notice anything unusual while talking with Erika?"
 - 9 Takeshi: "While talking with Erika, I felt like someone was approaching from the direction of the cliff. That's why I ended up running away."
 - 10 Masato: "Takeshi, how far away was this presence you felt?"
-

FIGURE 8
Example conversation in EQUAL condition.

-
- 1 Takeshi: "Where was everyone before Erika arrived?"
 - 2 Masato: "I was in the lounge with Ms.Iori looking for a ring from 20:45 to 22:00."
 - 3 Kozue: "I saw Yukiko saying she was going to the lounge at 21:15, what were you doing?"
 - 4 Kozue: "Were you in the lounge at that time, Yukiko?"
 - 5 Kozue: "Takeshi, did anyone hear you arguing with Erika?"
 - 6 Kozue: "Yukiko, have you heard anything about the argument with Erika?"
 - 7 Kozue: "Takeshi, what were you doing and where were you around 21:45 yesterday?"
 - 8 Takeshi: "Around 21:45, I was at the observation deck. I was talking with Erika."
 - 9 Kozue: "Takeshi, what were you talking about with Erika?"
 - 10 Kozue: "Takeshi, what were you talking about with Erika?"
-

FIGURE 9
Example conversation in SS condition.

and EQUAL conditions for coherence ($p = 0.084$), but significant differences between all other condition pairs ($p < 0.01$). For cooperativeness and diversity, significant differences were found between all condition pairs ($p < 0.01$).

Supplementary Figure S3 shows the results of the dialogue breakdown analysis conducted with Claude-3.7-sonnet and Gemini-1.5-pro. The judgments by Claude and Gemini were generally less strict compared to GPT-4, but the outcomes of Dunn's multiple comparisons were almost identical to those obtained using GPT-4. Supplementary Figure S4 presents the results of the LLM-as-a-judge using Claude and Gemini. From the Dunn's multiple comparisons, there do exist some pairs that differ in the statistical significance. However, there were

no instances where the order of medians was reversed with a statistical significance. To sum up, we did not observe a meaningful difference in the dialogue breakdown analysis and LLM-as-a-judge among the three LLMs.

Finally, we present the results of the human evaluation. For this evaluation, we recruited three external annotators who had substantial experience with Murder Mystery games (at least several years); two of them had experience serving as Game Masters. Annotators were presented 150 conversation sets (50 sets $\times$ 3 conditions) in random order and evaluated each utterance according to the criteria described in Section 4.2.3. Scores assigned to utterances were aggregated within each conversation set. First, we assessed inter-annotator

- 1

Kozue: "Did anyone see anything in the lounge last night?"
- 2

Masato: "No, I was searching for a ring in the lounge, but didn't see anything else in particular."
- 3

Kozue: "I see, thank you. Yukiko, you said you were going to the lounge around 21:15, did you see anything then?"
- 4

Yukiko: "No, I didn't meet anyone during that time. I saw Kozue leaving the building on my way to the lounge, but nothing else in particular."
- 5

Kozue: "Takeshi, do you remember where you were around 21:30 last night?"
- 6

Takeshi: "Kozue, I was at the observation deck around 21:30. I was waiting to talk with Erika. But Erika came much later."
- 7

Kozue: "I see, Takeshi. So Erika came after that. Around 21:15, I saw Erika arguing with someone, do you know who that was?"
- 8

Takeshi: "Ah, that was me. I was talking with Erika, but it turned into an argument. But I left right after that."
- 9

Kozue: "Someone might have approached Erika after Takeshi left... Masato, where were you during that time?"
- 10

Masato: "From 21:30 until after 22:00, I was searching for a ring in the lounge with Ms.Iori. We saw Yukiko going outside while we were searching, but what were you doing after that?"

FIGURE 10
Example conversation in CSSN-or-SS condition.

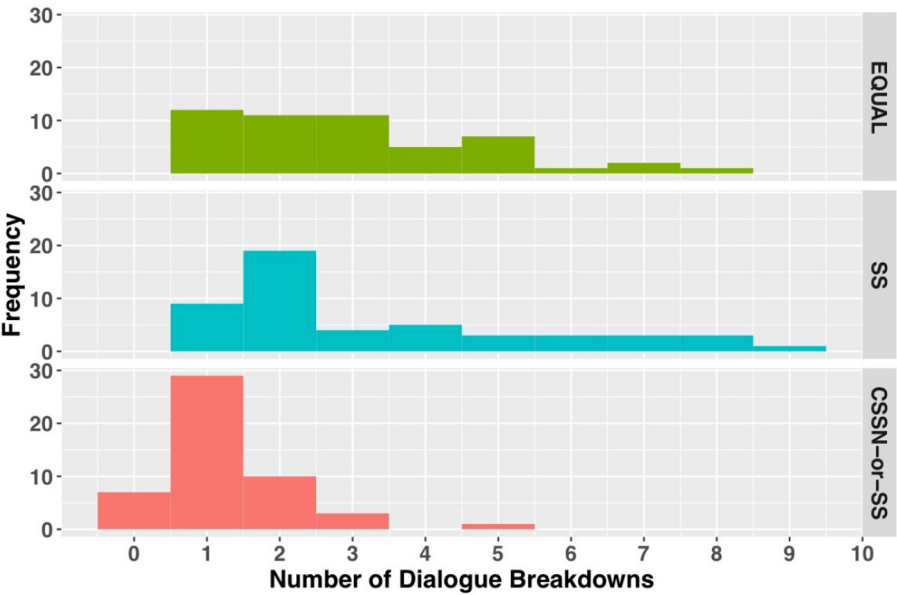
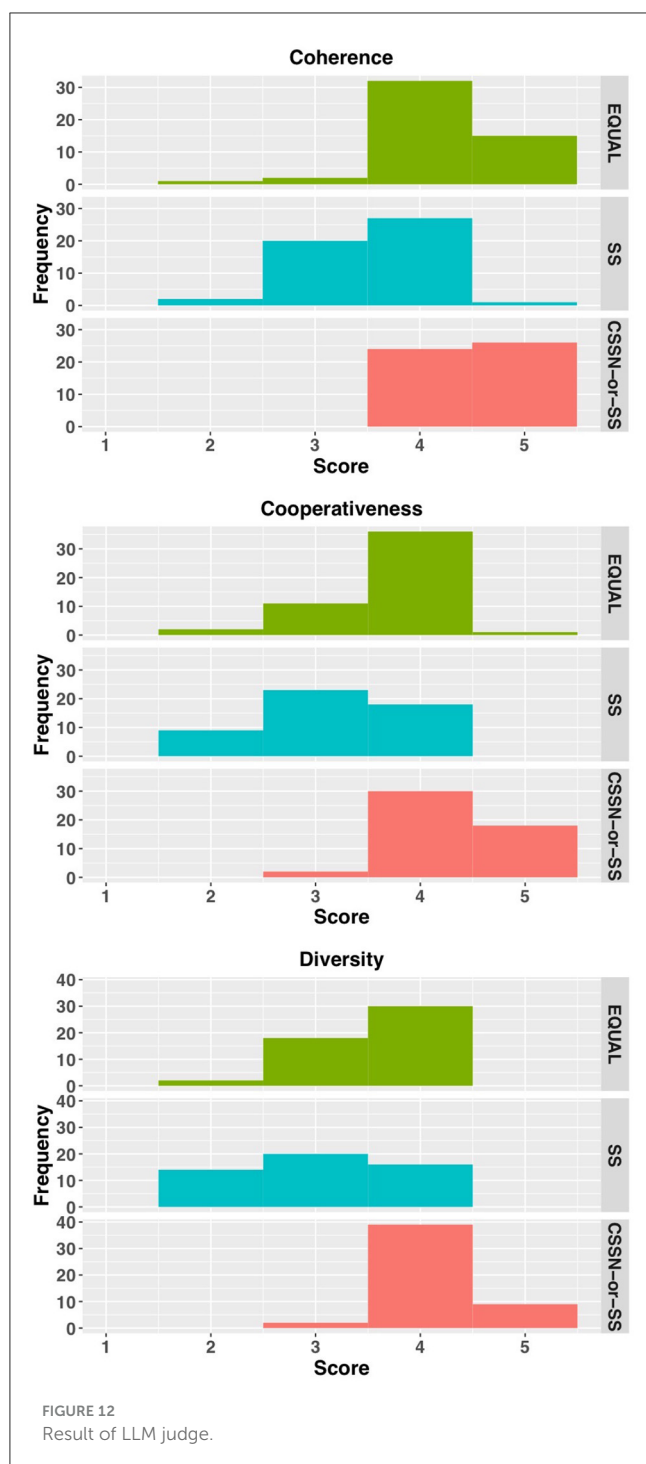


FIGURE 11
The number of utterances that lead to dialogue breakdowns within 10 turns.

agreement. For each conversation set, we calculated the agreement among annotators on binary judgments (positive or negative) for each evaluation item using Fleiss' kappa, resulting in $\kappa = 0.77$, which indicates substantial agreement. Hereafter, the scores averaged across annotators are used as the human judgement scores for each conversation set. The mean scores for the EQUAL, SS, and CSSN-or-SS

conditions were 5.89 (SD = 3.07), 4.61 (SD = 2.08), and 7.45 (SD = 4.13), respectively. A Kruskal-Wallis test revealed significant differences among conditions ($\chi^2 = 17.57$, $p < 0.001$). Figure 13 shows a histogram of the scores. Dunn's multiple comparison test (with Bonferroni correction) revealed that the CSSN-or-SS condition showed significantly higher scores compared to the SS condition ($p < 0.001$).



No significant differences were found between other pairs of conditions.

6 Discussion

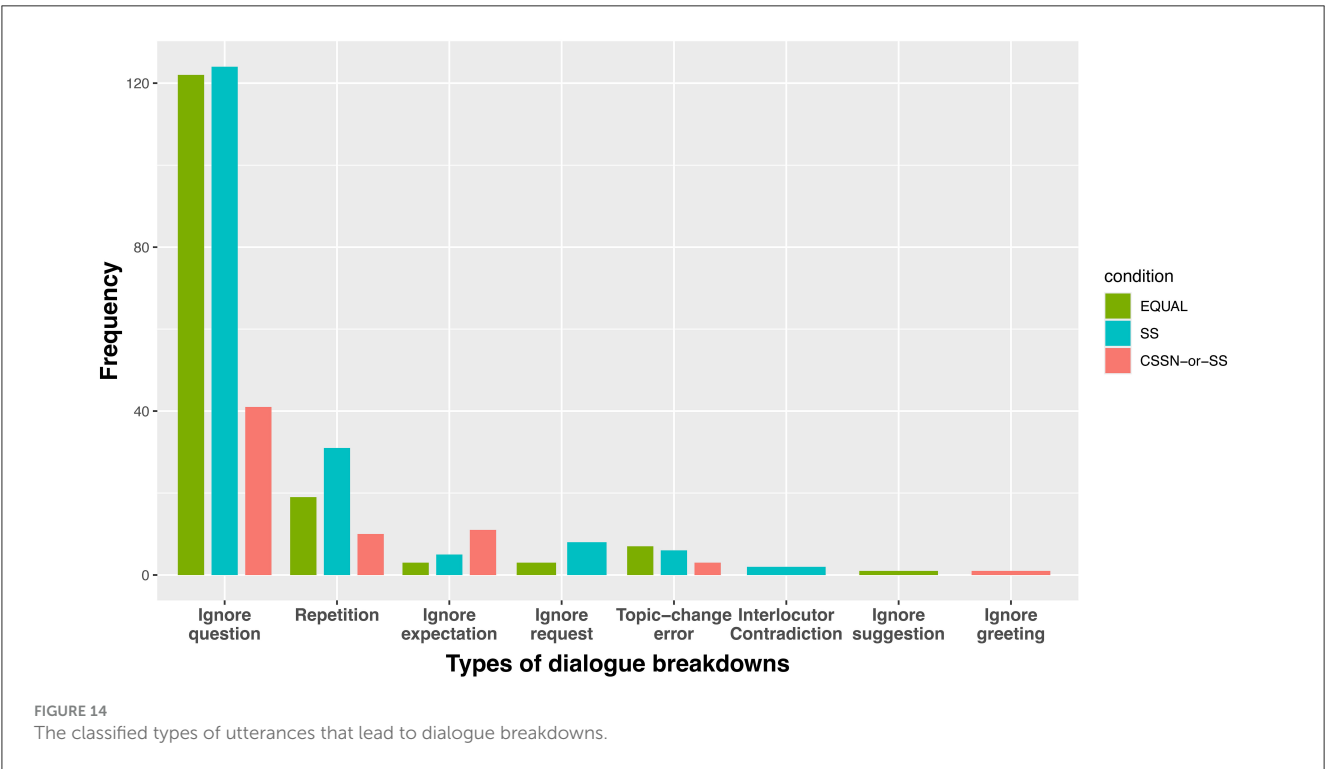
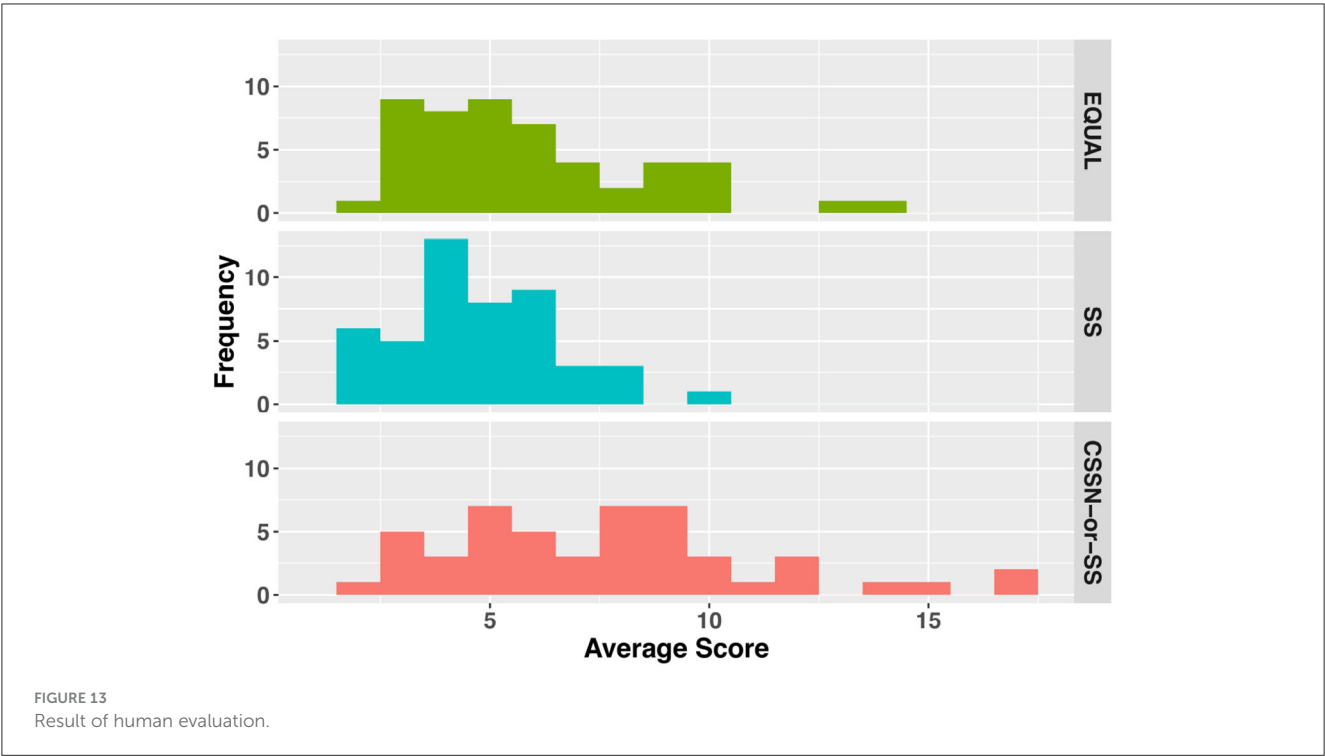
The experimental results of this study clearly demonstrate that the next-speaker selection mechanism utilizing adjacency pairs in turn-taking systems improves the quality of multi-party conversations in multiple aspects. From the analysis of

dialogue breakdowns, a significant decrease in the number of utterances that led to dialogue breakdowns was observed. Figure 14 shows the frequency distribution of classified dialogue breakdown types (refer to Supplementary Table S1) under each condition. In the CSSN-or-SS condition, a notable decrease in ignoring the question was confirmed compared to both the EQUAL and SS conditions. This is considered to be due to the next-speaker selection mechanism clarifying response obligations for specific participants, thereby suppressing inappropriate next speaker and responses to questions. Additionally, it is suggested that by structuring the flow of dialogue and promoting responses related to previous utterances, the Current Speaker Selects Next mechanism reduced abrupt topic changes (Topic-change error) and repetition. While approximately 40 instances of ignoring the question were identified in the CSSN-or-SS condition, detailed analysis of their content revealed characteristic patterns in addition to typical ignoring the question (e.g., cases where an agent with a response obligation asks a new question without answering). First, a tendency was observed where responses addressed only part of the question while avoiding core information. For example, as shown in Example 1 in Figure 15 where Kozue asked “Do you know anything about what Erika might have been hiding?”, Takeshi explained the circumstances of interaction with Erika but avoided addressing the essential answer about what was being hidden.

Second, some patterns were observed where agents intentionally shifted to different topics to avoid expected responses. As shown in Example 2 in Figure 15, despite expectations for discussion about Erika’s lipstick, Yukiko suddenly switched to discussing rings, representing a case of avoiding responding to the original question.

These characteristics suggest that within the context of reasoning games like Murder Mystery, the Current Speaker Selects Next mechanism influences agents’ information disclosure strategies. It is considered that as response obligations became clearer, agents began to control information disclosure in a more sophisticated way while avoiding simple ignoring the question to maintain their position in the game. For instance, in Example 2, as Erika’s lipstick was given as information that Yukiko needed to keep secret in her character settings, the switch to the topic of rings can be interpreted as a strategic choice to protect this secret. However, it is necessary to consider the possibility that these observed behavioral patterns might be influenced by the limitations in contextual processing capabilities of the LLMs used.

The evaluation results from the LLM-as-a-Judge demonstrate that the proposed method incorporating the Current Speaker Selects Next mechanism with adjacency pairs (CSSN-or-SS condition) comprehensively improved conversation coherence, cooperation, and diversity compared to both the EQUAL and SS conditions. These improvements can be attributed to the following advantages of introducing adjacency pairs: First, the generation of appropriate responses to questions was promoted, enabling logical conversation development. Second, clear turn-taking encouraged active participation in information exchange and problem-solving. Third, the repetition of identical utterances was suppressed, enabling the presentation of opinions from diverse perspectives.



However, it is noteworthy that no significant difference was observed between the CSSN-or-SS condition and EQUAL condition in terms of coherence evaluation. This result may be attributed to the characteristic properties of the EQUAL condition. Specifically, in the EQUAL condition, speaking opportunities are equally distributed among all conversation participants regardless of conversation content. Consequently, even when immediate response to the first part of an adjacency pair (e.g., question) is not

possible in the subsequent turn, participants are guaranteed to have a speaking opportunity in later turns, enabling them to provide the second part (e.g., answer). This structural characteristic may have ensured the eventual establishment of logical conversations, albeit not immediately.

As a supplementary analysis, we also evaluated dialogue on a per-turn basis, calculating the average scores for each metric across all turns. The results are shown in [Supplementary Figure S5](#). The

Example1

- 1 Kozue: "Yes, Masato. Around 21:45 last night, I saw Erika arguing with someone at the observation deck. I couldn't tell who it was, but I'm concerned about what Erika said. Takeshi, do you know anything about what Erika might have been hiding?"
- 2 Takeshi: "About what Erika was hiding... Actually, I met and talked with Erika at the observation deck around 21:30. But at that time, she didn't seem to be carrying anything."

Example2

- 1 Takeshi: "Well, Erika had been threatening us all this time. With that photo and all... I've been keeping an eye on Erika's behavior. Come to think of it, that lipstick, if Erika had it, there might be something important hidden there."
 - 2 Yukiko: "Maybe so. But first, we need to think about this ring. When I dropped it yesterday, it seems to have been switched – it's not the ring Takeshi gave me anymore. I'm worried ... "
-

FIGURE 15
Example of ignore expectation.

Kruskal-Wallis test and Dunn's multiple comparisons test revealed that, consistent with the whole dialogue evaluation, the CSSN-or-SS condition achieved significantly higher scores than the SS and EQUAL conditions. Notably, coherence scores were found to be abnormally high across all conditions when evaluated on a per-turn basis. This phenomenon can be attributed to the fact that turn-level evaluations focus primarily on the internal logic of individual utterances, while overlooking broader discourse-level aspects such as global consistency and inter-utterance coherence. Furthermore, diversity scores were consistently lower in the turn-level evaluation. We interpret this as a limitation of evaluating diversity from isolated turns, where broader patterns of opinion shifts and the variety of perspectives across a conversation are not adequately captured.

The results of the human evaluation revealed that the proposed method incorporating the Current Speaker Selects Next mechanism (CSSN-or-SS condition) yielded the highest mean score. This suggests that more information could be shared and used as a basis for reasoning in this condition. This interpretation aligns with the results of the LLM-based evaluations discussed above. Contrary to our initial expectations, the mean score for the EQUAL condition was comparatively high, approaching that of the CSSN-or-SS condition. In the EQUAL condition, the addressed agent appeared capable of reliably producing the second pair part of adjacency pairs, regardless of how long he or she had to wait until his or her turn came around, occasionally resulting in particularly high scores exceeding 10 points.

The presence of a certain number of low scores between 4 and 6 points even in the CSSN-or-SS condition indicates that there is still room for improvement in the proposed method. Analysis of low-scoring conversation examples, as shown in [Supplementary Figure S6](#), revealed a characteristic where excessive time was spent on specific topics. Specifically, in this conversation

example, 6 out of 10 turns were spent on speculations about the lipstick found in Yukiko's bag, yet they did not reach the truth about the lipstick (worth 3 points). The conversation ended without sharing or reasoning about other important information, resulting in a low score. While this conversation maintains a natural flow as general dialogue, it should be improved from the perspective of "sharing information and developing reasoning based on information," which is crucial in Murder Mystery games.

These analysis results suggest the importance of goal-oriented topic control. Specifically, the introduction of a mechanism that adjusts topic duration based on the importance of provided information could enable more effective reasoning processes.

Additionally, insufficient conversation turns might be one factor contributing to low scores. With fewer turns, discussions risk becoming biased toward specific topics, ending before other important information can be shared. In fact, by dedicating considerable time to speculations about the lipstick, other facts were neither shared nor verified, leaving the reasoning incomplete. To improve such situations, increasing conversation turns could potentially broaden the scope of discussion and promote the sharing and verification of crucial information related to the core of the case.

7 Conclusion

In this study, we implemented and verified the effectiveness of turn-taking systems, such as adjacency pairs discovered in conversation analysis research, in multi-party conversations among LLM-based agents. Based on Schegloff's theory that "in conversational turn-taking systems, the organization of utterance sequences, such as adjacency pairs, is the source of conversational coherence" (Schegloff, 1990), we aimed to achieve more natural

and coherent conversations by applying these norms to interactions between AI agents.

The experimental results strongly supported this theoretical prediction. The introduction of a turn-taking system using response obligations to the first pair part of an adjacency pair significantly reduced dialogue breakdowns, improved conversational cooperation and diversity, and enhanced agents' information sharing capabilities and reasoning abilities. In particular, the next-speaker selection mechanism based on adjacency pairs enabled smooth transitions of utterances between agents and promoted the generation of contextually appropriate responses. These results demonstrate that the norms of speech communication observed in human conversations also play a crucial role in conversations between AI agents.

These findings are not limited to the Murder Mystery game; they underscore the broader utility of multi-party agent conversations and have significant implications for the development of more capable and versatile multi-agent AI systems. The ability to manage turn-taking smoothly and maintain conversational coherence in multi-party interactions is a fundamental requirement for a wide array of potential applications. For instance, in collaborative AI systems, agents could work together to analyze complex data or solve engineering problems through structured discussion. In advanced simulation environments, such as modeling organizational dynamics or market behaviors, sophisticated multi-agent dialogue is needed to capture realistic interactions. Our work provides foundational insights and a proof-of-concept demonstrating that explicitly modeling conversational norms like turn-taking is a promising avenue for developing such systems applicable across these diverse domains.

However, several challenges remain in this research. The current system faces difficulties in maintaining memory using longTermMemory in extended dialogues of around 30 turns, and exhibits issues with topic management between agents, leading to topic deviation. Specific examples and detailed conversation logs are available on the project's [website](#).

It should be noted that our current simulation focuses exclusively on the discussion phase of Murder Mystery games, which represents only one critical component of the complete gameplay experience. A full game typically consists of multiple phases:

- Evidence-gathering phases (where players collect case-relevant information).
- Private negotiation phases (for strategic one-on-one interactions).
- Final accusation phases (where players vote on the murderer).

While the discussion phase serves as the core platform for information exchange and reasoning, simulating these additional phases would be essential for achieving comprehensive gameplay simulation. However, such an extension falls beyond the scope of this paper, which specifically investigates turn-taking mechanics in multi-party discussion contexts. While our evaluations focused on the quality of the dialogue process,

we acknowledge the lack of direct task success metrics, such as the accuracy or efficiency of identifying the murderer. Measuring such outcomes would provide a more complete picture of the system's effectiveness in the context of the game's objectives. Future work should integrate these additional game phases and incorporate task-based evaluations to fully evaluate how conversation quality impacts end-to-end game outcomes.

Also, it's important to acknowledge that adjacency pairs represent only one facet of the complex phenomenon of conversational coherence. Focusing solely on adjacency pairs can foreground immediate exchanges such as question-answer sequences, while potentially overlooking broader discourse-level relationships that span across longer stretches of conversation. Frameworks like Discourse Representation Theory (DRT) and Segmented Discourse Representation Theory (SDRT) (Kamp and Reyle, 1993; Asher et al., 2003) demonstrate that coherence involves not just local linking of turns, but also the maintenance of referents and semantic relationships across multiple sentences. These elements, while not fully captured by the adjacency pair model, are crucial for achieving coherence not only at a local level but also in terms of the overall goals and collaborative nature of communication.

Furthermore, future challenges include implementing a concept of time in conversation, such as the gradual prediction of transition-relevance places (TRPs) (Sacks et al., 1974) and controlling barge-in at non-TRPs, particularly in cases where listeners seek clarification, request additional explanation, raise questions, or express counterarguments during ongoing utterances.

Our current work on turn-taking provides a foundation for more natural AI agent conversations. However, we recognize that evaluating truly human-like behavior requires exploring a broader range of conversational phenomena beyond turn-taking. Therefore, a more comprehensive assessment of human-like interaction remains a key direction for future research.

Moving forward, we will address these challenges and further explore the applicability of conversation analysis theory in dialogues between AI agents. In particular, based on insights gained from the analysis of conversation data, we plan to improve the long-term memory mechanism and refine topic management.

Data availability statement

The original contributions presented in the study are included in the article/[Supplementary material](#), further inquiries can be directed to the corresponding author.

Author contributions

RN: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft. HM: Project administration, Resources, Supervision, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Acknowledgments

A part of this research was conducted as part of NINJAL Collaborative Research Project, “A Comprehensive Study of Spoken Language Using a Multi-Generational Corpus of Japanese Conversation.” A preprint version of this manuscript was previously deposited on arXiv (Nonomura and Mori, 2024) and is cited in the reference list. The authors acknowledge the preprint server for enabling early disseminations of this work in accordance with the journal’s preprint policy.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

- Anthropic (2024). *Claude 3.5 Sonnet*. Available online at: <https://www.anthropic.com/news/claude-3-5-sonnet> (accessed February 17, 2025).
- Asher, N., Asher, N. M., and Lascarides, A. (2003). *Logics of Conversation*. Cambridge: Cambridge University Press.
- Atkinson, R. C., and Shiffrin, R. M. (1968). “Human memory: a proposed system and its control processes,” in *The Psychology of Learning and Motivation* (New York: Academic Press), 89–195. doi: 10.1016/S0079-7421(08)60422-3
- Baddeley, A. D., and Hitch, G. (1974). “Working memory,” in *Psychology of Learning and Motivation* (New York: Academic Press), 47–89. doi: 10.1016/S0079-7421(08)60452-1
- Bailis, S., Friedhoff, J., and Chen, F. (2024). Werewolf Arena: A case study in LLM evaluation via social deduction. *arXiv [Preprint]* arXiv:2407.13943. doi: 10.48550/arXiv.2407.13943
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). “Language models are few-shot learners,” in *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS ’20* (Red Hook, NY: Curran Associates Inc), 1–25.
- Chiang, C.-H., and Lee, H. (2023). “Can large language models be an alternative to human evaluations?” in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, eds. A. Rogers, J. Boyd-Graber, and N. Okazaki (Toronto: Association for Computational Linguistics), 15607–15631.
- Clark, H. H., and Brennan, S. E. (1991). “Grounding in communication,” in *Perspectives on Socially Shared Cognition*, eds. L. B. Resnick, J. M. Levine, and S. D. Teasley (Washington, D.C.: American Psychological Association), 127–149.
- Clark, H. H., and Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition* 22, 1–39. doi: 10.1016/0010-0277(86)90010-7
- Filippas, A., Horton, J. J., and Manning, B. S. (2024). “Large language models as simulated economic agents: What can we learn from homo silicus?” in *Proceedings of the 25th ACM Conference on Economics and Computation, EC ’24* (New York, NY: Association for Computing Machinery), 614–615.
- Gao, M., Ruan, J., Sun, R., Yin, X., Yang, S., and Wan, X. (2023). Human-like summarization evaluation with ChatGPT. *arXiv [Preprint]* arXiv:2304.02554. doi: 10.48550/arXiv.2304.02554
- Gu, Z., Zhu, X., Guo, H., Zhang, L., Cai, Y., Shen, H., et al. (2024). AgentGroupChat: an interactive group chat simulacra for better eliciting emergent behavior. *arXiv [Preprint]*. Available at: <https://arxiv.org/abs/2403.13433>
- Herring, S. C. (1999). “Interactional coherence in CMC,” in *HICSS-32: Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences* (Los Alamitos, CA: IEEE Computer Society Press), 1–13. doi: 10.1109/HICSS.1999.772674
- Higashinaka, R., Araki, M., Tsukahara, H., and Mizukami, M. (2022). Classification of utterances that lead to dialogue breakdowns in chat-oriented dialogue systems. *J. Nat. Lang. Proc.* 29, 443–466. doi: 10.5715/jnlp.29.443
- Hu, B., Zheng, L., Zhu, J., Ding, L., Wang, Y., and Gu, X. (2024). Teaching plan generation and evaluation with GPT-4: unleashing the potential of LLM in instructional design. *IEEE Trans. Learn. Technol.* 17, 1471–1485. doi: 10.1109/TLT.2024.3384765
- Hu, S., Huang, T., Ilhan, F., Tekin, S., Liu, G., Kompella, R., et al. (2024). A survey on large language model-based game agents. *arXiv [Preprint]* arXiv:2404.02039. doi: 10.48550/arXiv.2404.02039
- Hugo, T., Martin, L., Stone, K., Albert, P., Almahari, A., Babaei, Y., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv [Preprint]* arXiv:2307.09288. doi: 10.48550/arXiv.2307.09288
- Jeon, J., and Lee, S. (2023). Large language models in education: a focus on the complementary relationship between human teachers and ChatGPT. *Educ. Inform. Technol.* 28, 15873–15892. doi: 10.1007/s10639-023-11834-1
- Junprung, E. (2023). Exploring the intersection of large language models and agent-based modeling via prompt engineering. *arXiv [Preprint]* arXiv:2308.07411. doi: 10.48550/arXiv.2308.07411
- Kamp, H., and Reyle, U. (1993). *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Dordrecht: Kluwer.
- Kocmi, T., and Federmann, C. (2023). “Large language models are state-of-the-art evaluators of translation quality,” in *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, eds. M. Nurminen, J. Brenner, M. Koponen, S. Latomaa, M. Mikhailov, F. Schierl, et al. (Allschwil: European Association for Machine Translation), 193–203.
- Lan, Y., Hu, Z., Wang, L., Wang, Y., Ye, D., Zhao, P., et al. (2023). LLM-based agent society investigation: Collaboration and confrontation in Avalon gameplay. *arXiv [Preprint]* arXiv:2310.14985. doi: 10.48550/arXiv.2310.14985
- Li, G., Al Kader Hammoud, H. A., Itani, H., Khizbullin, D., and Ghanem, B. (2023). “CAMEL: communicative agents for “mind” exploration of large language model

Generative AI statement

The author(s) declare that Gen AI was used in the creation of this manuscript. The authors declare that this manuscript includes text edited and translated with the assistance of gen AI tools, specifically Claude and DeepL. The final content has been reviewed and approved by the authors.

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/frai.2025.1582287/full#supplementary-material>

society,” in *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23* (Red Hook, NY: Curran Associates Inc).

Li, Z., Xu, X., Shen, T., Xu, C., Gu, J.-C., Lai, Y., et al. (2024). “Leveraging large language models for NLG evaluation: Advances and challenges,” in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y. N. Chen (Stroudsburg: Association for Computational Linguistics), 16028–16045.

Lu, K., Yu, B., Zhou, C., and Zhou, J. (2024). “Large language models are superpositions of all characters: Attaining arbitrary role-play via self-alignment,” in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, eds. L. W. Ku, A. Martins, and V. Srikumar (Bangkok: Association for Computational Linguistics), 7828–7840.

Meta Fundamental AI Research Diplomacy Team (FAIR), Bakhtin, A., Brown, N., Dinan, E., Farina, G., Flaherty, C., et al. (2022). Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science* 378, 1067–1074. doi: 10.1126/science.ade9097

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychol. Rev.* 63, 81–97. doi: 10.1037/h0043158

Modarressi, A., Imani, A., Fayyaz, M., and Schze, H. (2023). RET-LLM: Towards a general read-write memory for large language models. *arXiv*. [Preprint] arXiv:2305.14322. doi: 10.48550/arXiv.2305.14322

Nonomura, R., and Mori, H. (2024). Who speaks next? Multi-party AI discussion leveraging the systematics of turn-taking in murder mystery games. *arXiv* [Preprint] arXiv:2412.04937. doi: 10.48550/arXiv.2412.04937

OpenAI. (2023). GPT-4 technical report. *arXiv* [Preprint] arXiv:2303.08774. doi: 10.48550/arXiv.2303.08774

OpenAI (2022). *Introducing ChatGPT*. Available online at: <https://openai.com/index/chatgpt/> (accessed February 17, 2025).

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. (2023). “Generative agents: Interactive simulacra of human behavior,” in *UIST 2023: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (New York, NY: Association for Computing Machinery, Inc), 1–22.

Qian, C., Liu, W., Liu, H., Chen, N., Dang, Y., Li, J., et al. (2024). “ChatDev: Communicative agents for software development,” in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, eds. L. W. Ku, A. Martins, and V. Srikumar (Bangkok: Association for Computational Linguistics).

Rome, S., Chen, T., Tang, R., Zhou, L., and Ture, F. (2024). “Ask me anything”: How comcast uses LLMs to assist agents in real time,” in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York: ACM), 2827–2831.

Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language* 50, 696–735. doi: 10.1353/lan.1974.0010

Schegloff, E. A. (1968). Sequencing in conversational openings. *American Anthropologist* 70, 1075–1095. doi: 10.1525/aa.1968.70.6.02a00030

Schegloff, E. A. (1990). “On the organization of sequences as a source of ‘coherence’ in talk-in-interaction,” in *Conversational Organization and its Development*, ed. B. Dorval (Norwood, NJ: Ablex), 51–77.

Schegloff, E. A. (2007). *Sequence Organization in Interaction: A Primer in Conversation Analysis*. Cambridge: Cambridge University Press.

Schegloff, E. A., and Sacks, H. (1973). Opening up closings. *Semiotica* 8, 289–327. doi: 10.1515/semi.1973.8.4.289

Shanahan, M., McDonell, K., and Reynolds, L. (2023). Role play with large language models. *Nature* 623, 493–498. doi: 10.1038/s41586-023-06647-8

Shao, Y., Li, L., Dai, J., and Qiu, X. (2023). “Character-LLM: A trainable agent for role-playing,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, eds. H. Bouamor, J. Pino, K. and Bali (Singapore: Association for Computational Linguistics), 13153–13187.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., et al. (2023). LLaMA: Open and efficient foundation language models. *arXiv* [Preprint] arXiv:2302.13971. 10.48550/arXiv.2302.13971

Tulving, E. (1972). “Episodic and semantic memory,” in *Organization of Memory*, E. Tulving, and W. Donaldson (New York: Academic Press), 381–403.

Wang, N., Peng, Z., Que, H., Liu, J., Zhou, W., Wu, Y., et al. (2024). “RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models,” in *Findings of the Association for Computational Linguistics: ACL 2024* (Bangkok: Association for Computational Linguistics), 14743–14777.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., et al. (2022). “Chain-of-thought prompting elicits reasoning in large language models,” in *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22* (Red Hook, NY: Curran Associates Inc).

Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E. E., et al. (2024). “Autogen: enabling next-gen llm applications via multi-agent conversation,” in *COLM 2024*. University of Pennsylvania, Philadelphia, PA: Conference on Language Modeling. Available online at: <https://www.microsoft.com/en-us/research/publication/autogen-enabling-next-gen-llm-applications-via-multi-agent-conversation-framework/>

Yang, R., Chen, J., Zhang, Y., Yuan, S., Chen, A., Richardson, K., et al. (2024). SelfGoal: Your language agents already know how to achieve high-level goals. *arXiv* [Preprint] arXiv:2406.04784. doi: 10.48550/arXiv.2406.04784

Yao, S., Zhao, J., Yu, D., Du, N., Shafraan, I., Narasimhan, K., et al. (2022). ReAct: Synergizing reasoning and acting in language models. *arXiv* [Preprint] arXiv:2210.03629. doi: 10.48550/arXiv.2210.03629

Zhang, Z., Bo, X., Ma, C., Li, R., Chen, X., Dai, Q., et al. (2024a). A survey on the memory mechanism of large language model based agents. *arXiv* [Preprint] arXiv:2404.13501. doi: 10.48550/arXiv.2404.13501

Zhang, Z., Zhang-Li, D., Yu, J., Gong, L., Zhou, J., Liu, Z., et al. (2024b). Simulating classroom education with LLM-empowered agents. *arXiv* [Preprint] arXiv:2406.19226. doi: 10.48550/arXiv.2406.19226

Zhong, W., Guo, L., Gao, Q., Ye, H., and Wang, Y. (2024). “MemoryBank: enhancing large language models with long-term memory,” in *Proceedings of the 38th AAAI Conference on Artificial Intelligence* (Washington, DC: AAAI Press), 19724–19731.