Analysing Next Speaker Prediction in Multi-Party Conversation Using Multimodal Large Language Models

Taiga Mori, Koji Inoue, Divesh Lala, Keiko Ochi, Tatsuya Kawahara

Graduate School of Informatics, Kyoto University, Japan

Correspondence: mori.taiga.5h@kyoto-u.ac.jp

Abstract

This study analyses how state-of-the-art multimodal large language models (MLLMs) can predict the next speaker in multi-party conversations. Through experimental and qualitative analyses, we found that MLLMs are able to infer a plausible next speaker based solely on linguistic context and their internalized knowledge. However, even in cases where the next speaker is not uniquely determined, MLLMs exhibit a bias toward overpredicting a single participant as the next speaker. We further showed that this bias can be mitigated by explicitly providing knowledge of turn-taking rules. In addition, we observed that visual input can sometimes contribute to more accurate predictions, while in other cases it leads to erroneous judgments. Overall, however, no clear effect of visual input was observed.

1 Introduction

In recent years, research on dyadic conversation has made remarkable progress. With the advent of large language models (LLMs), it has become possible to generate coherent and contextually appropriate responses, enabling systems to engage in natural and practical human–AI dialogues. These advances have significantly enhanced conversational dynamics and language generation in two-party interactions.

However, when the number of participants increases, multi-party conversation still presents many challenges. Among them, turn-taking, the process by which speakers coordinate who talks next, is a fundamental yet difficult problem. Unlike dyadic dialogues, multi-party settings require the model to infer complex social and multimodal cues, such as gaze direction, body orientation, and addressing behaviours, to identify the next speaker correctly. Failure to predict turn transitions often leads to overlapping speech, unnatural pauses, or incoherent conversational flow.

In this study, we analyse how state-of-the-art multimodal large language models (MLLMs) can predict the next speaker in multi-party conversations (three participants). Specifically, the aim of this study is to clarify three points: how model size affects prediction accuracy, whether MLLMs can predict the next speaker without explicit knowledge of turn-taking rules, and whether visual information improves the accuracy of next speaker prediction.

The findings of this study not only reveal the current capabilities of MLLMs in understanding conversational dynamics, but also bridge classical theories and modern technologies, offering insights into how large foundation models can effectively participate in multi-party conversations. Accordingly, the contributions of this study are as follows.

- To investigate the extent to which state-of-the-art MLLMs can predict turn-taking behavior.
- To examine the effect of model size on turn-taking prediction performance.
- To determine whether explicit knowledge of turn-taking or visual information improves prediction accuracy.

2 Related Work

2.1 Turn-taking rules

Turn-taking is a phenomenon that humans perform naturally in everyday conversation, but it was first systematically modelled by Sacks et al. (1974). According to their framework, turn-taking occurs at transition-relevant places (TRP) through the recursive application of the following three rules (Sacks et al., 1974, p. 704):

> (a) If the turn-so-far is so constructed as
> to involve the use of a 'current speaker
> selects next' technique, then the party so
> selected has the right and is obliged to

Proceedings of the 16th International Workshop on Spoken Dialogue Systems Technology, pages 83–94
February 26–March 1, 2026. ©2026 Association for Computational Linguistics

> take next turn to speak; no others have such rights or obligations, and transfer occurs at that place.
>
> (b) If the turn-so-far is so constructed as not to involve the use of a ’current speaker selects next’ technique, then self-selection for next speakership may, but need not, be instituted; first starter acquires rights to a turn, and transfer occurs at that place.
>
> (c) If the turn-so-far is so constructed as not to involve the use of a ’current speaker selects next’ technique, then current speaker may, but need not continue, unless another self-selects.

These rules clarify that the next speaker can be predicted only when Rule (a), current speaker selects next, is applied. More precisely, prediction is feasible only when a single participant is explicitly selected as the next speaker; if multiple or unspecified recipients are addressed, determining who will speak next becomes inherently difficult. Therefore, when an (M)LLM attempts next speaker prediction, it should not only identify the most likely next speaker but also decide when no prediction can be made based on the available conversational cues.

Regarding the current speaker selects next technique, Sacks et al. (1974, p. 717) note that:

> Thus an important general technique whereby current speaker selects next—perhaps the central one—involves the affiliation of an address term (or some other device for achieving ’addressing’, e.g. gaze direction) to a first pair-part.

Thus, selecting the next speaker requires two key components: (1) addressing a specific participant within one’s utterance, and (2) performing an action that conditionally invites a response, such as producing the first pair-part of an adjacency pair. As Sacks and colleagues observed, addressing is inherently multimodal, involving linguistic and non-linguistic cues such as address terms and gaze direction. This implies that next speaker prediction cannot be achieved solely from linguistic information, a consideration that motivates the multimodal approach adopted in the present study.

2.2 Cues and models for next speaker prediction

Research on turn-taking has a long history, beginning with early studies that examined the relationship between gaze behaviour and the conversational roles.

Kendon (1967) analysed dyadic interactions recorded on film and found that eye gaze plays a central role in coordinating speaker changes. Specifically, speakers tend to look at their interlocutor at the end of their utterance to signal readiness for a role exchange, while listeners typically look away when beginning to speak, signalling acceptance of the turn. Kendon also observed that speakers’ gaze direction shifts depending on speech fluency, looking toward the listener during fluent passages and away during hesitations, suggesting that gaze functions both to control attention and manage turn-taking timing.

Building on this line of work, Jokinen et al. (2013) investigated the role of eye gaze in multi-party interactions, focusing on how gaze combines with speech features to predict turn transitions. Using eye-tracking data from casual three-party conversations, they trained Support Vector Machine (SVM) models on both gaze and speech features. Their results confirmed that eye gaze significantly contributes to predicting turn-taking activity, and that the speaker plays a particularly important role in coordinating the interaction space.

Beyond observational studies, several works have incorporated turn-taking mechanisms into interactive systems. Skantze et al. (2015) developed a robot dialogue system that engaged in collaborative card-sorting games with two human participants. Their system integrated multimodal cues, including speech, head pose, and object manipulation, to manage attention and turn-taking in a noisy, real-world environment. Their analysis of over 370 interactions showed how the system’s multimodal signals (e.g., filled pauses, gaze, and facial gestures) could help maintain smooth conversations despite some processing delays.

More recently, computational studies have applied machine learning and deep learning approaches to multi-party turn-taking prediction. Lee et al. (2023) proposed a Transformer-based model trained on high-fidelity multimodal data (gaze, head, body, and speech) from three-party interactions. Their model achieved over 80% accuracy in predicting turn transitions, and they introduced a new metric, Relative Engagement Level (REL) to quantify engagement in conversational units. Interestingly, they found that interlocutor state patterns (speaking, backchanneling, silence) were more predictive than gaze behaviour alone.

In a follow-up study, Lee and Deng (2024) addressed end-of-turn prediction in multi-party settings using a hybrid model that combined DistilBERT and a Gated Recurrent Unit (GRU). Their real-time window-based method outperformed traditional inter-pausal unit (IPU) approaches, effectively handling overlaps and interruptions that frequently occur in spontaneous multi-party dialogue. These results demonstrate the potential of pre-trained language models (PLMs) for improving the naturalness and flow of spoken dialogue systems.

Elmers et al. (2025) extended voice activity projection (VAP) models, previously used in dyadic dialogue, to triadic conversation in Japanese. Their models predicted each speaker's upcoming voice activity using only acoustic information, achieving higher accuracy than baseline models and showing that conversation type affects prediction performance.

More recently, researchers have begun exploring LLMs for next speaker prediction in multi-party dialogue. Inoue et al. (2025) examined whether GPT-4o, when prompted with transcripts of triadic discussions, could infer who would speak next. They reported that the model's accuracy was only slightly above chance, suggesting that LLMs still struggle to capture the complex contextual cues underlying next speaker selection. Their findings highlight the challenge of applying LLMs directly to conversation-level tasks without explicit modelling of turn-taking structure or multimodal information.

These studies suggest that turn-taking is a complex phenomenon involving not only linguistic information but also multimodal cues such as gaze, facial expressions, and prosody. Inoue et al. (2025) provide gaze information to LLMs in textual form. However, by supplying such information directly as images to MLLMs, the models themselves may be able to select salient cues, potentially leading to improved prediction accuracy. In addition, their study relies on relatively simple prompts, whereas explicitly incorporating knowledge of turn-taking rules into the prompt may further enhance performance. From a practical perspective, smaller models are also preferable due to their faster inference speed. Motivated by these considerations, the present study employs MLLMs to analyse how model size, the presence or absence of explicit turn-taking rules, and the inclusion of visual information affect next speaker prediction.

3 Dataset

This study used the Teidan corpus (Inoue et al., 2025). The corpus contains open-domain dialogues among triads of acquaintances on a variety of topics, such as "If Japan were to relocate its capital, where would it be?" (city), "If you could bring only one item to a deserted island, what would it be?" (island), "Where would you go if you were to travel this week?" (travel), "For a day off, would you go to the sea, mountains, or city?" (outdoor), "What is the most important thing in life?" (life), and "How would you travel from Tokyo to Osaka?" (trans).

Participants sat equidistantly around a round table, and the conversations were recorded using a camera placed in front of each participant and individual pin microphones worn by each speaker. A screenshot from the video recordings is shown in Figure 1.

This corpus is annotated with turn information. It originally includes manually transcribed IPU–level transcripts. Turns are constructed by merging IPUs in which a single speaker speaks continuously, excluding backchannels and laughter. These annotations were created by crowd workers.

We used data from 12 groups in the Teidan corpus. Each group participated in three sessions with different topics, and each session lasted approximately 5–7 minutes. The total duration of the data is 3 hours, 38 minutes, and 27 seconds, with a total of 3,121 turns. The average dialogue length is 6 minutes and 4 seconds, and the average number of turns per dialogue is 86.7. The duration and number of turns for each session are summarized in Appendix A. Note that A, B, and C refer to the same individual within a session but correspond to different individuals across sessions.

Figure 1: A snapshot from TEIDAN corpus

4 Next Speaker Annotation

As described in Section 2.1, next speaker prediction is possible only when the current speaker selects the next speaker, which makes it inappropriate to use the actual next speaker as the ground truth. Therefore, we manually annotated the next speaker for this study.

As described in Section 3, the corpus contains 3,121 turns. Due to time constraints, however, 1,000 samples were randomly selected for the experiments reported in this paper. The final turn of each conversation, which has no subsequent speaker, was removed in advance.

For these 1,000 samples, next speaker annotations were performed by the first author, who is trained in conversation analysis, based on the turn-taking rules described in Section 2.1. Specifically, when the current speaker addressed a particular listener —by referring to their name, gaze, or gesture— and produced the first-pair part of an adjacency pair, the turn was annotated with the label of the addressed listener (A, B, or C) as the next speaker. In all other cases, where no specific listener was expected to speak next, the label O (Other) was assigned. As a result of the annotation, the numbers of instances for A, B, C, and O were 53, 28, 46, and 873, respectively.

Enomoto et al. (2020) report that 21.2% of turn transitions involve current speaker selection, while 78.8% involve listener self-selection. In our annotation scheme, these correspond to labels A/B/C and O, respectively. In our dataset, O accounts for 87.3% of the labels, which is higher than that reported in their study, likely because our conversations are group discussions oriented toward reaching a shared conclusion, resulting in many utterances addressed to the group rather than to a specific individual.

5 Experiments

5.1 Experimental setup

In this study, we manipulated three factors: model size (three levels), presence or absence of turn-taking rules (two levels), and presence or absence of images (two levels), resulting in a total of 12 experimental conditions.

As state-of-the-art MLLMs, we used GPT-5, GPT-5-mini, and GPT-5-nano. The GPT-5 series, released by OpenAI in August 2025, is a multimodal foundation model capable of integrating multiple modalities such as text and images (OpenAI, 2025). It is also designed as a "reasoning model," which allows it to internally construct reasoning processes and generate logically consistent responses without requiring explicit step-by-step reasoning instructions in the prompt.

Hereinafter, we denote each experimental condition as *Model–Rule–Image*, where Model indicates the MLLM variant (G5, G5M, G5N), Rule indicates whether the turn-taking rules are provided (R) or not (NR), and Image indicates whether image input is used (I) or not (NI). For example, the condition using GPT-5 with the turn-taking rules and image input is denoted as G5-R-I.

We accessed the GPT-5 series through the Python API. The temperature parameter was fixed at 1 for all models by OpenAI, so no further tuning was applied. The model outputs were defined as dictionary objects containing both the predicted next speaker and the reasoning behind the prediction, specified via the *response_format* argument. All other parameters were kept at their default values.

The inputs were provided as the user role. The prompt was passed as text, and the images were encoded in Base64 format and passed as *image_url*. Each image was a screenshot of the video at the point when the turn ended (the same as in Figure 1), with a resolution of 1920×480.

Prompts we used in this paper consisted of three parts. The actual prompts are provided in the Appendix B. Note that in practice, the prompts were written in Japanese to match the language of the data.

The first part (enclosed by the orange box in Appendix B) described the basic task, provided the full dialogue history up to the current turn, and specified the response format. As discussed in Section 2.1, the next speaker can generally be predicted only when the current speaker explicitly selects the next speaker. If no specific participant was selected, any participant could speak next. Therefore, the task was defined as follows: if the next speaker could be determined, the model should respond with the participant label A, B, or C; if the next speaker could not be determined, it should respond with O. In addition, the model was asked to provide a concise reason for its prediction in no more than 30 words.

The second part (enclosed by the blue box in Appendix B) was included only when images were provided. The model was instructed to consider the relative positions of participants as well as gaze direction, gestures, and body orientation in its reasoning.

The third part (enclosed by the green box in Appendix B) was used only when the turn-taking rules of Sacks et al. (1974) was explicitly provided. The model was guided to first estimate the addressee and the dialogue act in the current turn, with specific examples provided based on Kadota et al. (2024) and Iseki et al. (2019). If a participant was addressed and the dialogue act was of a type that expected a specific response, akin to the first part of an adjacency pair, the addressed participant was assigned as the next speaker (A, B, or C). Otherwise, if no such condition applied, the model was instructed to respond with O, indicating that the next speaker could not be determined.

5.2 Evaluation metrics

We evaluated each condition using the following three metrics. First, when considering the application of our method to a dialogue system, the system corresponds to one of the participants A, B, or C. In this setting, if the model predicts that "any participant may speak next" even though a specific participant has actually been selected as the next speaker, the system may interrupt another participant's turn. This situation is the most critical one to avoid. Therefore, among the cases in which a specific participant is actually selected as the next speaker, we define the proportion of instances in which the model predicts a specific participant as the next speaker as the *Interruption Avoidance Rate (IAR)*. In interaction settings where interruptions are not acceptable, methods with a high IAR are desirable.

$$\mathrm{IAR} = \frac{|\{i \mid y_i \in S \land \hat{y}_i \in S\}|}{|\{i \mid y_i \in S\}|} \tag{1}$$

Here, let $y_i \in \{A, B, C, O\}$ denote the ground-truth label of the i-th instance, and let $\hat{y}_i \in \{A, B, C, O\}$ denote the corresponding predicted label. Let N be the total number of instances. We define the set of speaker labels as $S = \{A, B, C\}$. Note that this metric does not take into account whether the model correctly predicts which of A, B, or C is the next speaker, that is, it does not consider the model's ability to discriminate among individual speakers.

Next, to evaluate this discriminative ability, we use the macro-averaged F1 score computed from the precision and recall of each of the labels A, B, and C, considering only cases in which a specific participant is selected as the next speaker.

$$\begin{aligned}\mathrm{Precision}_s &= \frac{|\{i \mid y_i = s \land \hat{y}_i = s\}|}{|\{i \mid y_i \in S \land \hat{y}_i = s\}|}\\ \mathrm{Recall}_s &= \frac{|\{i \mid y_i = s \land \hat{y}_i = s\}|}{|\{i \mid y_i = s\}|}\\ \mathrm{F1}_s &= \frac{2 \cdot \mathrm{Precision}_s \cdot \mathrm{Recall}_s}{\mathrm{Precision}_s + \mathrm{Recall}_s}\\ \text{Macro-F1} &= \frac{1}{|S|}\sum_{s \in S} \mathrm{F1}_s\end{aligned} \tag{2}$$

Here, a lowercase symbol $s \in S$ denotes a specific human speaker.

Finally, when no specific participant is selected as the next speaker, but the model predicts that a specific participant is the next speaker, no turn-taking problem arises if that participant is the system itself. However, if the predicted participant is another participant, the system loses its opportunity to speak. In data such as ours, where instances labelled as O are frequent, this can become a critical issue. Therefore, among the cases in which no specific participant is selected as the next speaker, we define the proportion of instances in which the model predicts that no specific participant is selected as the *Speaking Opportunity Detection Rate (SODR)*. In scenarios where the system is expected to actively participate in the conversation, methods with a high SODR are desirable.

$$\mathrm{SODR} = \frac{|\{i \mid y_i = O \land \hat{y}_i = O\}|}{|\{i \mid y_i = O\}|} \tag{3}$$

5.3 Results

Figures 2–3 show the experimental results, and the detailed numerical values are provided in Appendix C.

Figure 2 shows the scores of the three-evaluation metrics under each experimental condition. First, with respect to the macro-F1 score, the lowest value was observed under the G5N-R-NI condition at 55%, while the highest value was obtained under the G5-R-I condition at 93%.

Next, regarding the IAR, the lowest score was 57% under the G5N-R-NI condition. The highest score was 100%, achieved under both the G5M-NR-I and G5-NR-I conditions.

Finally, for the SODR, the lowest value was observed under the G5-NR-I condition at 0.1%,

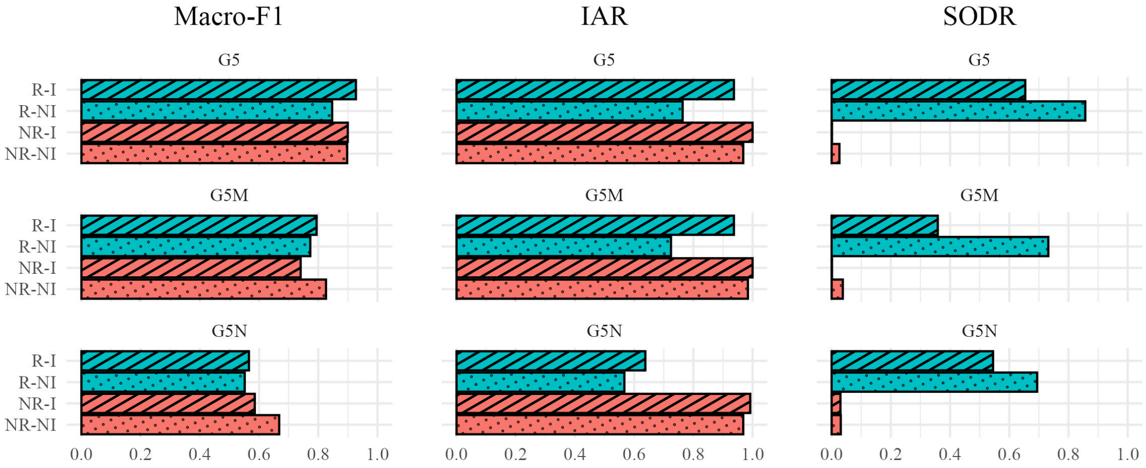
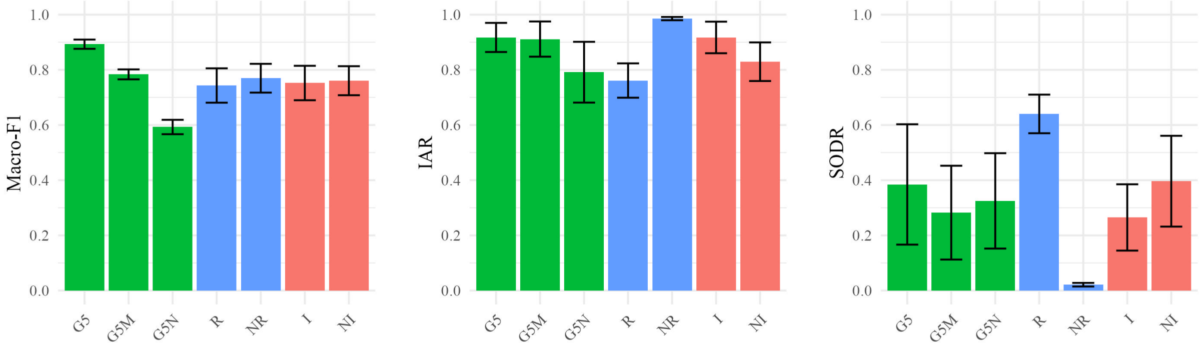


Figure 2: Evaluation scores for each condition
Blue indicates the presence of turn-taking rules, and hatched bars indicate the use of image input.



(a) Macro-F1

(b) IAR

(c) SODR

Figure 3: Average macro-F1, IAR, and SODR at each variable level

whereas the highest value was 86% under the G5-R-NI condition.

Figure 3a shows the average macro-F1 scores for each level of the experimental variables. The error bars represent the standard error. As shown in the figure, the presence or absence of turn-taking rules and visual input does not affect the discriminative performance of next speaker prediction, whereas larger models achieve higher discriminative performance.

Figure 3b shows the average IAR for each level of the experimental variables. The figure indicates that neither model size nor the presence of visual input results in substantial differences in IAR, however, IAR is higher when the turn-taking rules are not applied than when they are applied. Specifically, the average IAR is 99% when the turn-taking rules are not applied, whereas it decreases to 76% when the turn-taking rules are applied.

Figure 3c shows the average SODR for each level of the experimental variables. Similar to IAR, no substantial differences are observed with respect to model size or the presence of visual input. In contrast to IAR, however, SODR is markedly higher when the turn-taking rules are applied than when they are not applied. Specifically, the average SODR is 64% when the turn-taking rules are applied, whereas it decreases to 2% when the turn-taking rules are not applied.

5.4 Discussions

With respect to the macro-F1 score, we found that neither the presence of turn-taking rules nor visual input has a significant effect, whereas performance increases proportionally with model size. This result indicates that, when the next speaker is selected by the current speaker, the models are able to predict the next speaker using linguistic context and their internal knowledge alone. Furthermore, the prediction performance improves as the model size increases.

In contrast, for IAR and SODR, model size and the presence of visual input do not have a notable impact, whereas the presence or absence of turn-taking rules has a clear effect. Specifically, IAR is higher when the turn-taking rules are not applied, while SODR is higher when the turn-taking rules are applied. This pattern arises because these two metrics are in a trade-off relationship by definition. When the turn-taking rules are not included in the prompt, the proportion of predictions labelled as O becomes markedly smaller, which leads to an increase in IAR but a decrease in SODR. In fact, under conditions with the turn-taking rules, label O accounts for 59% of the predictions on average, whereas under conditions without the turn-taking rules, it accounts for only 2% on average. Although the relative importance of IAR and SODR depends on the conversational setting and the relationships among speakers, in the conversational scenarios examined in this study, the absence of turn-taking rules results in the system being almost unable to utilize its speaking opportunities.

Furthermore, the finding that the proportion of predictions labelled as O becomes markedly smaller than the actual proportion in the absence of turn-taking rules indicates that, when the models are not explicitly provided with knowledge of turn-taking rules, they tend to exhibit a bias toward overpredicting a single next speaker, even in situations where no specific next speaker is actually determined.

Finally, the presence or absence of images did not affect any of the results. Three possible reasons can be considered. First, the models may be able to predict the next speaker in most cases solely from linguistic context. Second, because still images cannot capture dynamic cues such as gaze shifts or bodily movements, the visual information may have been insufficient. Third, images contain a large amount of information, which may have made it difficult for the models to attend to the cues that are truly relevant for next speaker prediction.

6 Case Analysis

Next, we conducted a qualitative analysis comparing the reasoning provided by the models in the experiment with what actually occurred in each case, in order to identify the patterns in the model's next speaker predictions.

The following example is taken from a discussion on the topic of relocating Japan's capital. The model was asked to predict the next speaker based on the information up to the final line.

(1) session-02-city_8
01 B: Okay, um, so then, uh, today's topic is what place could become the next capital if a massive earthquake hit directly under Tokyo?
02 C: That's a scary topic.
03 B: Yeah, pretty scary.
04 A: It is scary.
05 B: Well, then it's gotta be Osaka, right? It's the second-largest city by population, and, you know, Japan's second-biggest city should probably take over if something happens to the biggest one.
06 C: Well, if you think about it naïvely, yeah, that makes sense.
07 B: Naïvely.
08 C: Naïvely.
09 B: So, what do you think?

The utterance in line 09 is a question asking for an opinion, and therefore a response is expected in the following turn. Since at this point in the conversation, neither A nor C has expressed his own opinion yet, if the question is not directed toward a specific individual, either A or C could take the next turn.

In line 09, B does not linguistically address a specific person, therefore, under the G5-R-NI condition, the model judged that the question was open to anyone and predicted O. However, in the actual conversation, B directs his gaze toward A and also points at A with his hand. This clearly indicates that B is physically addressing A, and indeed, A becomes the next speaker. Therefore, if the models had a correct understanding of turn-taking mechanisms, the appropriate prediction should have been A.

Under the G5-R-I condition, the model responded "B's utterance is a question requiring a response. B was previously interacting with C, but in the image B's gaze is directed toward A. Since A has not yet spoken, it is likely that B is inviting A's opinion." In this case, the models successfully utilized the visual information in making an appropriate prediction.

In the following case, the participants are discussing what single item they would bring to a deserted island.

(2) session-01-island_40
01 C: (omitted) Right, even if you had a lighter, you'd have to keep the fire going once it's lit. If you run out of fuel, you can't use it anymore. When you live alone, you know, there's also that problem of… who's going to watch the fire?
02 A: Yeah, on a deserted island you're basically living alone in most cases.

In line 01, C expresses the opinion that even with a lighter, it would be difficult for one person to keep a fire burning. In line 02, A agrees with this opinion. Since A's utterance is not the first pair part of an adjacency pair, it is an utterance that allows any participant to take the next turn. Indeed, under the G5-R-NI condition, the model judged that "A's utterance is a casual comment that neither addresses a specific addressee nor functions as a question or request; therefore, it has low response relevance, and the next speaker cannot be determined," and correctly predicted O. In the actual conversation, the next speaker was B.

However, under the G5-R-I condition, the model reasoned "A responds to C's mention of 'living alone,' confirming agreement with the phrase '…dakara ne (that's why).' His gaze is directed toward C, suggesting that C's acknowledgment or response is expected," and predicted C. In this case, the addition of visual information led the model to make an incorrect prediction.

These two examples suggest that visual information can sometimes contribute to accurate predictions but can also lead to incorrect ones, which may explain why no clear overall trend was observed in the experiments.

In the final case, the participants are again discussing the topic of relocating Japan's capital city.

(3) session-01-city_3
01 B: Well, for example, I think somewhere in the Tokai region would be better.
02 A: Ah, I see.

In line 01, B expresses the opinion that the Tokai region would be preferable. In line 02, A shows understanding of that opinion. Since A's utterance is not the first pair part of an adjacency pair, the next speaker is not designated, and if the models properly understood the conversation, the appropriate response would be O. In fact, under both the G5-R-I and G5-R-NI conditions, the model judged that no specific response was expected and correctly predicted O. However, in the actual interaction, the next speaker was B.

By contrast, under the G5-NR-I condition, the model reasoned: "A's utterance 'Ah, I see' functions as a backchannel marking completion of the turn. C, who asked the previous question, has not yet expressed her own opinion, so it is natural for her to respond or elaborate next," and predicted C. Although the model correctly recognized that no specific response was expected, its prediction was incorrect, as the actual next speaker was B.

Interestingly, however, immediately after line 02, B says, "C, what do you think specifically?", inviting C to express her opinion. Thus, while B was in fact the next speaker, the model's prediction reflects an understanding consistent with the participants' own expectations that C should speak next in this context.

These findings suggest that even without explicit turn-taking rules, the models can often infer from context who would naturally speak next. However, to accurately predict actual turn transitions, explicit knowledge of turn-taking rules remains necessary.

7 Conclusions

This study has two main limitations. First, because we evaluated only the GPT-5 series as MLLMs, it remains unclear whether the findings of this study can be generalized to other MLLMs. Second, our experiments used only a single image captured near the end of each turn, which prevents the models from capturing dynamic information such as gaze shifts and timing.

Regarding future work, the first direction is to conduct experiments with other models, such as Gemini 2.5 (Comanici et al., 2025) and Qwen2-VL (Wang et al., 2024), to examine whether similar trends can be observed. The second direction is to investigate whether prediction performance can be improved by incorporating multiple consecutive frames as input to account for gaze dynamics, or, conversely, by reducing the input information—such as providing gaze information in textual form as in (Inoue et al., 2025) or masking irrelevant regions of the images.

Acknowledgments

This work was supported by JST Moonshot R&D JPMJPS2011 and JST PRESTO JPMJPR24I4.

References

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Mikey Elmers, Koji Inoue, Divesh Lala, and Tatsuya Kawahara. 2025. Triadic multi-party voice activity projection for turn-taking in spoken dialogue systems. *arXiv preprint arXiv:2507.07518*.

Mika Enomoto, Yasuharu Den, and Yuichi Ishimoto. 2020. A conversation-analytic annotation of turn-taking behavior in japanese multi-party conversation and its preliminary analysis. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 644–652.

Koji Inoue, Divesh Lala, Mikey Elmers, Keiko Ochi, and Tatsuya Kawahara. 2025. An llm benchmark for addressee recognition in multi-modal multi-party dialogue. *arXiv preprint arXiv:2501.16643*.

Yuriko Iseki, Keisuke Kadota, and Yasuharu Den. 2019. Characteristics of everyday conversation derived from the analysis of dialog act annotation. In *2019 22nd conference of the oriental cocosda international committee for the co-ordination and standardisation of speech databases and assessment techniques (o-cocosda)*, pages 1–6. IEEE.

Kristiina Jokinen, Hirohisa Furukawa, Masafumi Nishida, and Seiichi Yamamoto. 2013. Gaze and turn-taking behavior in casual conversational interactions. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 3(2):1–30.

Keisuke Kadota, Seima Oyama, and Yasuharu Den. 2024. Annotation of addressing behavior in multi-party conversation. In *2024 27th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–6. IEEE.

Adam Kendon. 1967. Some functions of gaze-direction in social interaction. *Acta psychologica*, 26:22–63.

Meng-Chen Lee and Zhigang Deng. 2024. Online multimodal end-of-turn prediction for three-party conversations. In *Proceedings of the 26th International Conference on Multimodal Interaction*, pages 57–65.

Meng-Chen Lee, Mai Trinh, and Zhigang Deng. 2023. Multimodal turn analysis and prediction for multi-party conversations. In *Proceedings of the 25th International Conference on Multimodal Interaction*, pages 436–444.

OpenAI. 2025. Introducing gpt-5. https://openai.com/ja-JP/index/introducing-gpt-5/. Accessed October 22, 2025.

Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *language*, 50(4):696–735.

Gabriel Skantze, Martin Johansson, and Jonas Beskow. 2015. Exploring turn-taking cues in multi-party human-robot discussions about objects. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 67–74.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

A Statistics of turn annotation

Session ID	Time (mm:ss)	Turn (A/B/C)	Session ID	Time (mm:ss)	Turn (A/B/C)
session-01-city	6:14	13/17/13	session-07-life	5:35	26/26/21
session-01-island	6:37	13/22/13	session-07-outdoor	5:44	27/23/15
session-01-travel	8:11	18/24/17	session-07-trans	5:53	42/44/38
session-02-city	5:51	16/22/24	session-08-life	6:56	15/15/12
session-02-island	6:22	28/41/30	session-08-outdoor	5:39	11/13/13
session-02-travel	5:18	22/34/32	session-08-trans	5:59	11/19/26
session-03-city	6:12	19/29/30	session-09-life	6:40	17/26/23
session-03-island	7:29	20/30/24	session-09-outdoor	5:30	16/35/40
session-03-travel	5:40	30/30/31	session-09-trans	5:40	15/40/49
session-04-city	5:46	33/22/27	session-10-life	5:38	44/31/46
session-04-island	5:51	37/22/27	session-10-outdoor	5:45	40/26/37
session-04-travel	5:40	26/18/19	session-10-trans	5:50	36/26/35
session-05-city	5:19	45/44/48	session-11-life	6:25	25/24/35
session-05-island	5:29	42/64/45	session-11-outdoor	6:36	44/34/46
session-05-travel	5:01	34/57/45	session-11-trans	5:48	37/26/41
session-06-city	6:13	19/16/25	session-12-life	6:26	20/16/13
session-06-island	6:17	48/37/63	session-12-outdoor	5:37	37/22/30
session-06-travel	6:13	41/14/43	session-12-trans	7:03	40/23/26
Average	6:04	86.7	Total	3:38:27	3121

B Prompts

```
# Task
For the following last utterance in a conversation among three people, A, B, and C, estimate which listener will be the next speaker.
Your answer must be one of "A", "B", "C", or "O".
"A", "B", "C": the respective participants
- "O": if it is not possible to identify a specific next speaker
# Output Format
Return your output only in the following JSON format. Do not include any extra text.
{
   "next_speaker": "A|B|C|O",
   "rationale": "Briefly explain your reasoning in 30 words or less"
}
# Scenario
A, B, and C are freely discussing topics such as "If you were to move the capital of Japan, where would it be?”, "If you could take only one thing to a desert island, what would it be?”, "Where would you go if you were going somewhere this week?”, "If you travel from Tokyo to Osaka, what would you use?” and "What is the most important thing in your life?"

    # Image information
    The image shows the situation at the moment the last utterance ends.
    - From right to left, A, B, and C are seated.
    - From A’s perspective, C is on the left and B is on the right.
    - From B’s perspective, A is on the left and C is on the right.
    - From C’s perspective, B is on the left and A is on the right.
    - Consider information such as gaze, gestures, and body orientation when making your judgment.

    # Reasoning Procedure
    **Step 1: Analyze the Addressee**
    Determine who the last utterance is addressed to, considering:
    - Whether it contains a specific participant’s name
    - Whether discourse markers like "but" or "so" follow another’s utterance
    - References to a participant, their utterance, experience, or knowledge
    - Responses, repairs, or co-constructions directed at someone
    - Polite language, gaze, gestures, or object use directed at a specific participant
    **Step 2: Analyze the Dialogue Act**
    Determine what kind of dialogue act the last utterance performs. If multiple acts occur, choose the one performed last. Examples:
    information provision / information request / confirmation request / response / request / instruction / command / invitation /
    suggestion / offer / acceptance / refusal / feedback
    If the act expects a response, such as a question or suggestion, the addressee is likely the next speaker.
    **Step 3: Estimate the Next Speaker**
    If the utterance is directed to a specific addressee and expects a response → the addressee is the next speaker.
    If the addressee is unspecified or addressed to everyone, or the act does not expect a response → the next speaker cannot be identified
    (O).

# Dialogue
A: ...
B: ...
```

C Detailed experimental results

Condition	Macro-F1	IAR	SODR
G5N-R-NI	0.55	0.57	0.69
G5N-R-I	0.57	0.64	0.55
G5N-NR-NI	0.67	0.97	0.03
G5N-NR-I	0.59	0.99	0.03
G5M-R-NI	0.77	0.72	0.73
G5M-R-I	0.79	0.94	0.36
G5M-NR-NI	0.83	0.99	0.03
G5M-NR-I	0.74	1.0	0.001
G5-R-NI	0.85	0.76	0.86
G5-R-I	0.93	0.94	0.65
G5-NR-NI	0.90	0.97	0.03
G5-NR-I	0.90	1.0	0.001

Variable level	Macro-F1		IAR		SODR	
	Mean	SE	Mean	SE	Mean	SE
G5	0.89	0.02	0.92	0.05	0.39	0.22
G5M	0.78	0.02	0.91	0.06	0.33	0.17
G5N	0.59	0.03	0.79	0.11	0.28	0.17
R	0.74	0.06	0.76	0.06	0.02	0.007
NR	0.77	0.05	0.99	0.005	0.64	0.07
I	0.75	0.06	0.83	0.07	0.27	0.12
NI	0.76	0.05	0.92	0.06	0.40	0.17