Analysis of the Correlation Between Theory of Mind and Dialogue Ability to Identify Essential ToM for Dialogue Systems

Haruhisa Iseno[1,2], Atsumoto Ohashi[1,2], Tetsuji Ogawa[3], Shinnosuke Takamichi[4], Ryuichiro Higashinaka[1,2]

[1]Graduate School of Informatics, Nagoya University, [2]NII LLMC,
[3]Department of Communications and Computer Engineering, Waseda University,
[4]Department of Information and Computer Science, Keio University

{iseno.haruhisa.h4@s.mail, ohashi.atsumoto.c0@s.mail,higashinaka@i}.nagoya-u.ac.jp

ogawa.tetsuji@waseda.jp, shinnosuke_takamichi@keio.jp

Abstract

In large language models (LLMs), improvements in theory of mind (ToM), which is the ability to infer others' mental states, are expected to enhance dialogue performance. However, the quantitative verification of this relationship remains insufficient. Therefore, this study evaluates the performance of seven high-performing LLMs across three ToM benchmarks (ToMBench, FANToM, and Hi-ToM) and six dialogue tasks to verify the correlation between ToM and dialogue performances. Our findings revealed a fundamental correlation between ToM and dialogue performance, though significant differences emerged depending on the ToM aspects examined. Specifically, we observed high correlations with dialogue performance for both ToM evaluated in conversational formats and ToM assessed with questions directly asking about beliefs. Additionally, ToM in situations involving conflicting beliefs between agents strongly correlates with dialogue performance. Furthermore, stable correlations were observed between first-order ToM and dialogue capabilities. These findings provide crucial guidelines for developing dialogue systems with human-like dialogue capabilities.

1 Introduction

Dialogue systems based on large language models (LLMs) have recently demonstrated remarkable performance improvements across diverse dialogue tasks (OpenAI et al., 2023; Yi et al., 2024). To achieve more human-like advanced dialogue capabilities, it is essential to enhance not only language processing but also the ability to understand and reason about the mental states of others, i.e., theory of mind (ToM).

Numerous benchmarks have been proposed to evaluate ToM in LLMs, as improvements in dialogue capabilities through enhanced ToM are believed to be vital (Le et al., 2019; Gandhi et al., 2023; Wu et al., 2023; Kim et al., 2023; Chen et al., 2024; Shinoda et al., 2025). These benchmarks evaluate ToM by presenting LLMs with stories or dialogue histories and conducting question-answering tasks regarding the mental states of the people involved, such as their beliefs and intentions. Despite such efforts being undertaken, the relationship between LLMs' ToM benchmark performance and dialogue performance has not been quantitatively verified, and it remains unclear whether performance improvements in ToM benchmarks lead to improvements in dialogue performance.

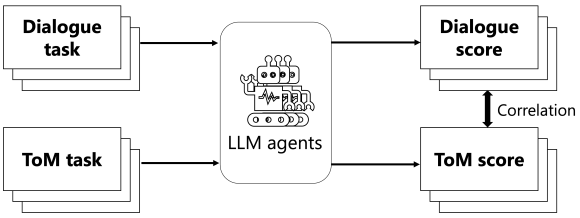

Figure 1: Experimental framework for analyzing correlations between dialogue and ToM task performances in LLMs.

Therefore, this study quantitatively examines the extent to which existing ToM benchmarks accurately capture the ToM required for dialogue. Specifically, we evaluated the performance of seven state-of-the-art LLMs on three different ToM benchmarks (ToMBench, FANToM, and Hi-ToM) and six dialogue tasks (Taboo, Wordle, Drawing, Reference Game, Private & Shared, and MutualFriends), and systematically investigated the correlation between ToM and dialogue performance.

The contributions of this study are as follows.

- We developed a framework for analyzing the relationship between LLMs' ToM and dialogue performance, and verified their relation-

ship through correlation analysis of multiple ToM benchmarks and dialogue tasks (Fig. 1).

- We found strong correlations between ToM performance, particularly when evaluated in conversational formats or through questions directly probing beliefs, and overall dialogue performance. Moreover, ToM in scenarios where others hold beliefs that differ from one's own showed a strong correlation with dialogue performance.

- We observed stable correlations between first-order ToM and dialogue capabilities; however, correlations decreased markedly for second- and higher-order ToM, suggesting that the current dialogue tasks may be insufficient for capturing the relevance of higher-order ToM abilities.

2 Related Work

In this section, we review the ToM benchmarks and recent evaluations of dialogue system performance.

2.1 ToM Benchmarks

A variety of benchmarks for evaluating ToM from diverse perspectives have been proposed (Le et al., 2019; Ma et al.; Gandhi et al., 2023; Wu et al., 2023; Chen et al., 2024; Xu et al., 2024). Most of these benchmarks measure the accuracy for problems regarding characters' mental states, such as beliefs and intentions, based on story contexts, as exemplified by the Sally-Anne task (Baron-Cohen et al., 1985). For example, ToMBench (Chen et al., 2024) comprehensively analyzes ToM using 20 types of diverse story-format problems.

ToM benchmarks using conversations as contexts have also been proposed. Benchmarks such as FANToM (Kim et al., 2023), NegotiationToM (Chan et al., 2024), and ToMATO (Shinoda et al., 2025) adopt ToM evaluation with conversations as the context and attempt to construct ToM evaluation environments that approximate the actual interaction settings.

Most benchmarks evaluate first-order ToM (estimating a person's mental state) and second-order ToM (estimating a person's mental state about another person's mental state). An exception is HiToM (Wu et al., 2023), which adopts a design that systematically evaluates higher-order ToM (up to the fourth order) in addition to the conventional first- and second-order ToM, enabling the measurement of more complex ToM.

Based on these benchmarks, ongoing discussions have focused on whether LLMs exhibit ToM. Kosinski (2023) reported that GPT-3.5 demonstrated a performance equivalent to children aged 7–10 years on ToM tasks, suggesting the possibility that ToM spontaneously emerged in LLMs. Conversely, Ullman (2023) and Shapira et al. (2024) showed that the accuracy rate of LLMs decreased significantly with only minor modifications to ToM benchmarks, arguing that LLMs have not developed ToM but rather solved ToM problems by relying on superficial pattern matching. In the current study, we assume that there is some relationship between LLMs' ToM and dialogue capabilities and investigate which aspects of ToM correlate and to what extent.

2.2 Evaluation of Dialogue System Performance

Dialogues are broadly classified as task-oriented and non-task-oriented (McTear, 2022). While the evaluation methods differ for each type, this study focuses on task-oriented dialogue, which allows for easier quantitative evaluation to conduct correlation analysis.

Benchmarks such as MultiWOZ (Budzianowski et al., 2018) and schema-guided dialogue (Rastogi et al., 2020) measure dialogue system performance by utilizing dialogue state tracking accuracy, response generation quality, and task success in specific tasks, such as restaurant reservations and hotel searches. Numerous collaborative dialogue tasks requiring conversational grounding for task completion have also been proposed (He et al., 2017; Udagawa and Aizawa, 2019; Kim et al., 2019; Bara et al., 2021). These tasks measure the ability to establish common ground (Clark, 1996) with interlocutors through multi-turn dialogues.

Frameworks for automatic evaluation of the dialogue capabilities of LLMs using dialogue between LLMs have recently been developed. Chalamalasetti et al. (2023) proposed Clembench, a framework that evaluates the dialogue capabilities of LLMs through LLM-to-LLM interactions in a game format, enabling a comprehensive automatic evaluation by quantitatively measuring LLM performance across multiple dialogue tasks.

In the current study, we focus on dialogue tasks that require conversational grounding, where ToM is strongly involved, and analyze the relationship

between dialogue performance and ToM. We use Clembench to evaluate the dialogue performance.

3 Approach

We conduct a correlation analysis between LLMs' ToM and dialogue performance to verify their relationship. Correlation analysis has frequently been utilized to examine whether evaluation metrics can appropriately measure the intended targets. For example, in machine translation, Papineni et al. (2002) demonstrated a strong correlation between BLEU scores and human evaluations. The confirmation of this correlation led to the understanding that improvements in BLEU scores directly lead to the generation of better-quality translations for humans.

Although BLEU has sometimes been used for dialogue evaluation, Liu et al. (2016) demonstrated that the correlation between human dialogue evaluation and BLEU-based dialogue evaluation is not significantly high. Therefore, other evaluation measures have increasingly been used for dialogue evaluation (Zhang et al., 2019; Mehri and Eskenazi, 2020).

Since evaluation metrics must appropriately measure what they are intended to, and ToM is regarded as a fundamental ability underpinning human social interaction (Baron-Cohen et al., 1985; Frith, 1994), it is important to examine whether the performance measured by ToM benchmarks is related to actual dialogue capabilities. Therefore, this study analyzes the correlation between ToM and dialogue capabilities. As shown in Fig. 1, we use m ToM benchmarks and l dialogue tasks on n state-of-the-art LLMs and calculate the correlation coefficients between LLMs' ToM benchmark accuracy rates and dialogue task scores. We verify whether correlations exist between ToM and dialogue capabilities by conducting a correlation analysis individually for each aspect of ToM and determine which aspects exhibit stronger correlations.

The following sections describe the selection of the ToM benchmarks and dialogue tasks addressed in this study.

3.1 Selection of ToM Benchmarks

Although evaluating the overall ToM performance is important, evaluating the performance across various aspects of ToM is necessary to comprehensively assess the correlations with dialogue performance. Therefore, this study selects benchmarks by focusing on the following aspects that can be particularly relevant to dialogue performance.

The first aspect is the context format. ToM benchmarks include tasks that estimate characters' beliefs from narrative-format contexts, such as the Sally-Anne task (Wimmer and Perner, 1983; Baron-Cohen et al., 1985), and tasks that infer characters' mental states from conversational texts. Even within narrative formats, there are settings that include dialogue between characters within the story and other settings that do not. To analyze the impact of these differences in context format on correlations with dialogue capabilities, we select benchmarks to comprehensively cover cases where contexts are in narrative and dialogue formats, and cases where narratives include dialogue and those that do not.

The second aspect is the mental state targeted for inference. In ToM tasks, ToM that infers belief states different from those of the reasoner (e.g., a situation where one knows that the cookies are in box B, but the other person who does not know this believes they are in box A) are important (Quesque and Rossetti, 2020; Shinoda et al., 2025), as in false belief tasks (Wimmer and Perner, 1983). However, it is unclear which ability—estimating mental states different from one's own or estimating the same mental states—is more strongly related to actual dialogue capabilities. Therefore, to investigate this relationship, we select benchmarks that include both types of inference.

The third aspect is the question format. ToM benchmarks include diverse question formats that directly ask about beliefs, such as "Where does A think X is?", and formats that require identifying knowledge holders, such as "Who knows where X is?" Although these questions necessitate different response content, correctly answering them requires the same ToM: "Where does a person think X is?" To investigate whether differences in question format affect correlations with dialogue capabilities even when the required ToM is the same, we select benchmarks that include different question formats.

The final aspect is the order of beliefs. First-order beliefs are a person's own beliefs, such as "A thinks X." Second-order beliefs are beliefs about others' beliefs, such as "B thinks that A thinks Y." Furthermore, some ToM tasks measure ToM for even higher-order beliefs, such as third- and fourth-order beliefs. To analyze the correlations between ToM for each of these first- to fourth-order beliefs

and dialogue capabilities, we select benchmarks that include ToM for beliefs of multiple orders.

3.2 Selection of Dialogue Tasks

To verify the correlation between ToM and dialogue performance, we select task-oriented dialogue tasks in which task achievement can be quantified with clear evaluation metrics and conversational grounding is required for task completion; such grounding is believed to be strongly related to ToM.

4 Experiments

We conducted experiments to investigate the correlation between ToM and dialogue performance. First, we selected ToM and dialogue tasks to be used in the correlation analysis. We then used seven state-of-the-art LLMs (GPT4.1, Gemini2.5-Flash, Claude4-Sonnet, Grok4, Llama3.3-70B, Qwen3-32B, and Mistral-Small) to answer ToM tasks and perform dialogue tasks. Subsequently, we examined the correlation between ToM and dialogue performance.

4.1 ToM Tasks

We selected three ToM benchmarks (ToMBench, FANToM, and Hi-ToM) that comprehensively encompass the differences in context format, mental states targeted for inference, question format, and order of beliefs, enabling a correlation analysis from each perspective. Examples of these benchmarks are provided in Appendix A, and their details are as follows.

ToMBench A benchmark that estimates characters' belief states using stories as context. A distinctive feature of ToMBench is its comprehensive evaluation of ToM in 20 diverse story contexts. These stories include various tasks, ranging from those based on the classic Sally-Anne task to mental state estimations in more complex social situations. Each story is structured with questions about the characters' mental states, enabling the measurement of whether LLMs can accurately estimate the characters' beliefs, desires, intentions, and other mental states based on narrative contexts. In this dataset, most problems involve first-order ToM, while one of the 20 stories, the false-belief task, involves ToM for both first- and second-order beliefs.

FANToM A benchmark that estimates specific speakers' belief states using multiparty conversations as context. A distinctive feature of this benchmark is that dynamic information asymmetry arises when characters leave and rejoin the dialogue. Since conversations continue even when speakers are absent, the known information differs among characters, resulting in a structure in which each character develops a different belief state. FANToM has three subtasks: (1) BeliefQA, which are tasks that directly ask about characters' belief states; (2) InfoAccessibilityQA (InfoQA), which are tasks that enumerate people who possess specific information; and (3) AnswerabilityQA (AnsQA), which are tasks that enumerate people who can correctly answer BeliefQA questions. BeliefQA includes ToM problems for first- and second-order beliefs and is classified into two conditions: accessible and inaccessible. In the accessible condition, questions are asked about belief states in which information is shared among characters, whereas in the inaccessible condition, questions are asked about belief states in which information is not shared.

Hi-ToM A benchmark that evaluates higher-order ToM. Its distinctive feature is requiring ToM up to the fourth order. It adopts narratives that extend the Sally-Anne task as context, with settings in which multiple characters enter and exit rooms while moving objects. Each story includes five questions that gradually increase in complexity, from "Where is X?" (0th-order) to "Where does A think B thinks C thinks D thinks X is?" (4th-order). Additionally, there are two types of settings: those that include communication between characters in the story (Tell condition) and those that do not (No_Tell condition).

We used the accuracy rates of the LLMs on these three ToM benchmarks as indicators of LLMs' ToM performance.

To analyze the effects of the differences in context formats, we utilized ToMBench and Hi-ToM as representative narrative-format benchmarks and FANToM as a representative dialogue-format benchmark. Furthermore, we analyzed the changes in the correlations caused by the presence or absence of communication elements in narrative-format benchmarks by comparing correlations under the two experimental conditions of Hi-ToM's Tell and No_Tell conditions.

To analyze the effects of the differences in mental states targeted for inference, we used FANToM's accessible and inaccessible conditions. The

Task	Task description	Evaluation metrics
Taboo	A vocabulary explanation task in which one of two players explains a target word without using specific forbidden words, while the other player guesses the target word from the explanation.	A score comprising the accuracy rate of finding target words and the number of dialogue turns until success.
Wordle	A deduction task in which players identify a five-letter English word within six attempts. After each attempt, feedback is provided indicating whether correct letters are in the correct positions.	A score comprising the accuracy rate of identifying five-letter English words and the number of dialogue turns until success.
Drawing	A task in which one of two players describes a virtual image composed of a 5×5 character grid using only language, while the other player reconstructs the original character string.	F1 score between the reconstructed character string and the original character string.
Reference Game	A reference resolution task in which two players identify one common image from among three virtual images composed of 5×5 character grids.	Accuracy rate of selecting the correct image.
Private & Shared	A task in which a questioner and answerer share information through dialogue, and the answerer estimates which information the questioner knows and does not know at each point in the dialogue.	An integrated score combining information sharing success rate and accuracy rate of estimating the partner's belief state.
MutualFriends	A task in which two speakers are each given different friend lists and find mutual friends through dialogue.	Accuracy rate of correctly identifying common friends.

Table 1: Task descriptions and evaluation metrics of dialogue tasks used in this study.

accessible condition requires the inference of mental states identical to that of the reasoner, whereas the inaccessible condition requires the inference of mental states that are different from those of the reasoner. This enables the analysis of changes in correlation caused by the differences in targeted mental states (inference of mental states identical to one's own versus inference of mental states different from one's own).

To analyze the effects of the differences in question formats, we used FANToM's three subtasks (BeliefQA, InfoAccessibilityQA, and AnswerabilityQA). As these questions ask for the same ToM inference through different question formats, we analyzed the effects of the differences in the question formats on the correlations.

To analyze the effects of the differences in order of beliefs, we utilized Hi-ToM's first- to fourth-order ToM tasks and first- and second-order belief estimation tasks from ToMBench and FANToM. We analyzed the effects of the differences in the order on the correlations by comparing the correlations for each order.

4.2 Dialogue Tasks

To evaluate the dialogue capabilities of LLMs, we implemented five types of text games conducted through dialogue (Taboo, Wordle, Drawing, Reference Game, Private & Shared). These five tasks are also used by Clembench (Chalamalasetti et al., 2023). In addition, we used the MutualFriends (He et al., 2017) task as the sixth dialogue task. All of these tasks have quantitative evaluation metrics and are tasks in which ToM is believed to be important for task completion. The description and evaluation metrics of each dialogue task are provided in Table 1.

All tasks other than Wordle were conducted through dialogue between the same LLMs. In contrast, since Wordle can proceed with only simple feedback from the user, dialogue tasks were conducted through dialogue between LLMs and a rule-based user simulator. Subsequently, based on these dialogues, the LLM performance in each dialogue task was scored on a 0–100 scale to indicate the dialogue performance. For MutualFriends, the scores were calculated using only dialogue success rates, and for all other tasks, the scoring method defined by Clembench (Chalamalasetti et al., 2023) was applied. Furthermore, the average score of the six tasks was calculated as "Average" and utilized to indicate the overall dialogue capabilities of the LLMs.

	ToMBench	FANToM	Hi-ToM
Drawing	0.52	**0.68**	0.15
Private & Shared	0.68	**0.76**	0.73
Reference Game	0.65	**0.91**	0.50
Taboo	0.69	**0.91**	0.61
Wordle	0.66	**0.89**	0.26
MutualFriends	0.37	**0.54**	0.31
Average	0.66	**0.85**	0.45

Table 2: Pearson correlation coefficients between ToM tasks and each dialogue task. Bold values indicate the strongest correlations for each dialogue task.

	No_Tell	Tell
Drawing	0.15	0.12
Private & Shared	0.72	0.61
Reference Game	0.43	0.52
Taboo	0.58	0.56
Wordle	0.29	0.16
MutualFriends	0.27	0.31
Average	0.43	0.39

Table 3: Pearson correlation coefficients under settings where narrative tasks include interactions (Tell) and settings where they do not (No_Tell).

4.3 Experiment Procedure

To execute the ToM tasks, we presented the LLMs with story or dialogue texts as context and asked them to answer questions about the characters' mental states in a multiple-choice format. For Hi-ToM, we used the existing question-answering prompts included in the dataset, while for ToMBench and FANToM, we designed new prompts for this study (see Appendix B).

To execute the dialogue tasks, we controlled the LLMs using the existing prompts provided by the benchmark when the five tasks included by Clembench (Taboo, Wordle, Drawing, Reference Game, Private & Shared) were performed. We used the prompts designed for this study when the MutualFriends task was performed (see Appendix C).

We quantitatively evaluated the ToM and dialogue performance of each model using the above-mentioned methods and calculated the Pearson correlation coefficients between them. Given our sample size of seven state-of-the-art LLMs, correlation coefficients of ≥ 0.670 indicate a significant trend ($p < 0.1$), ≥ 0.755 indicate statistical significance ($p < 0.05$), and ≥ 0.875 indicate high statistical significance ($p < 0.01$). However, with such a limited sample size, individual correlation coefficients may lack statistical robustness. Therefore, rather than relying solely on the statistical significance of individual correlations, we prioritized identifying consistent overall trends that emerged across multiple tasks and conditions, using the correlation coefficients as a reference for interpreting the strength and direction of observed relationships.

4.4 Results

This section presents the experimental results of the correlations between ToM and dialogue performance based on the four perspectives mentioned in Section 3.1: (1) context format, (2) mental states targeted for inference, (3) question format, and (4) order.

4.4.1 Effect of Context Type

Table 2 lists the correlation coefficients between the overall accuracy rates of the three ToM benchmarks (ToMBench, FANToM, and Hi-ToM) and the success rates of each dialogue task. The results show a clear trend in the correlations with dialogue tasks. In almost all dialogue tasks, FANToM, which uses conversations as context, exhibits higher correlations than the narrative-format ToMBench and Hi-ToM. For the Private & Shared task, the correlation coefficients are nearly equivalent for the three ToM benchmarks. This is likely because this task is originally designed to perform question-answering tasks with content similar to ToM tasks, resulting in a high structural similarity with ToM benchmarks.

Table 3 presents the results of a comparative analysis of the correlations with dialogue tasks in Hi-ToM tasks under settings where communication occurs between characters (Tell condition) and where communication does not occur (No_Tell condition). The results indicate no significant changes in the correlation caused by the presence or absence of communication elements. The results reveal fundamental limitations of narrative-format contexts. Specifically, even when introducing a small number of conversational elements within stories, the same improvement in the correlation with dialogue capabilities as when using conversations as the context is not observed. We therefore consider the inclusion of characters' utterances in stories insufficient to bring about essential improvements in dialogue capability.

4.4.2 Effect of Mental State Type

We analyzed the extent to which ToM toward others who hold belief states different from their own correlates with dialogue performance. Table 4 presents the results of classifying the BeliefQA task within

	Accessible	Inaccessible
Drawing	−0.74	**0.71**
Private & Shared	−0.85	**0.96**
Reference Game	−0.64	**0.82**
Taboo	−0.72	**0.88**
Wordle	−0.77	**0.82**
MutualFriends	−0.84	**0.75**
Average	−0.85	**0.91**

Table 4: Pearson correlation coefficients between each of the accessible and inaccessible conditions in FANToM and dialogue tasks. Bold values indicate higher correlations for each dialogue task.

Model	Accessible	Inaccessible
GPT4.1	80.26	72.10
Gemini2.5-Flash	79.71	78.05
Claude4-Sonnet	62.34	80.16
Grok4	61.61	80.06
Llama3.3-70B	77.15	62.54
Mistral-Small	92.14	28.90
Qwen3-32B	93.24	20.24

Table 5: Task accuracy rates for each LLM on FANToM.

the FANToM benchmark into accessible and inaccessible conditions and comparing the correlations with dialogue tasks. The results show that the inaccessible condition exhibits strong positive correlations across all dialogue tasks, whereas the accessible condition consistently exhibits strong negative correlations. This result indicates that the ability to estimate belief states that differ from one's own is important for high dialogue capabilities. The ability to correctly grasp situations in which others hold beliefs that are different from one's own and predict their actions and reactions on the basis of those beliefs is thus considered a crucial capability that determines success in an actual dialogue.

For a more detailed analysis of the causes of the negative correlations observed in the accessible condition, we individually compared each LLM performance under accessible/inaccessible conditions. Table 5 lists the accuracy rates of each LLM on the FANToM benchmark under accessible/inaccessible conditions. As we can see, Mistral and Qwen3 exhibit extremely high accuracy rates in the accessible condition, but low accuracy rates in the inaccessible condition. Models that scored highly in the accessible condition tend to answer assuming that all information is shared without considering others' perspectives and might have been conducting inference based on the incorrect assumption that "all information is equally accessible to all people." The negative correlation is also caused by Grok4 and Claude4, which performed well in the dialogue tasks but exhibited relatively low performance in the 60-point range in the accessible condition.

	BeliefQ	AnsQ	InfoQ
Drawing	<u>0.65</u>	0.41	**0.67**
Private & Shared	**0.94**	<u>0.45</u>	0.44
Reference Game	**0.83**	0.78	<u>0.81</u>
Taboo	**0.88**	<u>0.75</u>	<u>0.75</u>
Wordle	<u>0.78</u>	0.72	**0.86**
MutualFriends	**0.67**	0.26	<u>0.36</u>
Average	**0.87**	0.59	<u>0.72</u>

Table 6: Pearson correlation coefficients for each FANToM subtask. Bold values indicate the strongest correlations for each dialogue task, and underlined values represent the second strongest correlations.

4.4.3 Effect of Question Format

For the correlations with dialogue capabilities owing to differences in question format, we analyzed the relationship between FANToM's three subtasks (BeliefQ, AnsQ, and InfoQ) and dialogue task performance. As presented in Table 6, BeliefQ consistently exhibits higher correlations than the other two subtasks. These results clearly indicate that although all three subtasks of FANToM require the same ToM inference, the correlations with dialogue capabilities differ significantly depending on the question format. Although formats that ask for direct belief estimation, such as BeliefQ, exhibit high correlations with dialogue tasks, formats that ask for applications of inference results, such as AnsQ and InfoQ, exhibit low correlations despite dealing with the same ToM inference. This result indicates that, even for problems requiring the same ToM inference, the obtained evaluation varies depending on the question format. In particular, the results clearly show that problem formats that perform direct estimation of beliefs are more reliable indicators of dialogue capabilities.

4.4.4 Effect of Reasoning Order

To examine how the order of inference in ToM affects dialogue performance, we conducted a correlation analysis between performance by order of beliefs on three ToM benchmarks (Hi-ToM, ToMBench, FANToM) and dialogue task performance.

Tables 7 and 8 present the correlation coefficients between ToM up to the fourth order for each

	0th	1st	2nd	3rd	4th
Drawing	0.57	**0.83**	−0.33	−0.44	−0.41
Private & Shared	**0.99**	0.74	0.28	0.12	0.29
Reference Game	**0.80**	0.62	0.02	−0.03	0.15
Taboo	**0.86**	0.57	0.16	0.08	0.33
Wordle	**0.70**	0.66	−0.27	−0.38	−0.11
MutualFriends	**0.75**	0.72	−0.15	−0.28	−0.24
Average	**0.84**	0.80	−0.08	−0.20	−0.05

Table 7: Pearson correlation coefficients between ToM by order in Hi-ToM and dialogue tasks. Bold values indicate the strongest correlations for each dialogue task, and underlined values represent the second strongest correlations.

	ToMBench		FANToM	
	1st	2nd	1st	2nd
Drawing	**0.63**	−0.18	**0.73**	0.48
Private & Shared	**0.73**	0.18	**0.94**	0.88
Reference Game	**0.80**	−0.06	**0.83**	0.77
Taboo	**0.84**	0.15	**0.88**	0.83
Wordle	**0.81**	−0.00	**0.83**	0.66
MutualFriends	**0.44**	−0.07	**0.72**	0.54
Average	**0.78**	−0.02	**0.91**	0.75

Table 8: Pearson correlation coefficients between ToM by order in ToMBench and FANToM and dialogue tasks. Bold values indicate the strongest correlations for each dialogue task.

benchmark and dialogue task performance. The results show that the correlations with dialogue capabilities differ between lower-order ToM (0th- and 1st-order) and higher-order ToM. Across all benchmarks, the first-order ToM exhibits stable positive correlations with dialogue tasks. In particular, the high correlation with first-order belief estimation in FANToM indicates that first-order ToM evaluation in a conversational format is strongly related to dialogue capabilities. In contrast, second-order and higher-order ToM exhibit markedly low correlations, with negative or no correlations observed in many cases. In Hi-ToM, correlations decrease significantly from second-order ToM onward. In ToMBench and FANToM, second-order ToM also exhibits lower correlations compared to first-order ToM.

These results do not mean that second- and higher-order ToM are unnecessary in dialogue. Although both are important cognitive abilities in complex interactions, within the scope of the collaborative tasks addressed in this study, the recognition of world states (0th-order ToM) and estimation of others' beliefs and intentions (first-order ToM) likely played more important roles. In an actual human dialogue, different orders of ToM inference are dynamically utilized depending on the situation. To achieve fundamental improvements in the LLMs' ToM in the future, the introduction of more advanced tasks that require higher-order ToM inference (Wang et al., 2019; Kano et al., 2024) will be necessary.

5 Conclusion

This study conducted a comprehensive correlation analysis of the relationship between ToM and dialogue performance in LLMs using three ToM benchmarks and six dialogue tasks. We confirmed a fundamental correlation between ToM and dialogue performance, though significant differences emerged depending on the ToM aspects examined. Specifically, we observed high correlations with dialogue performance for both ToM evaluated in conversational formats and ToM assessed with questions directly asking about beliefs. Additionally, ToM in situations involving conflicting beliefs between agents strongly correlates with dialogue performance. Our findings indicate that dialogue tasks requiring higher-order ToM inferences are crucial for a more comprehensive evaluation of the dialogue capabilities of LLMs. This study provides the first systematic empirical analysis of the relationship between ToM and dialogue performance, with the results serving as valuable guidelines for developing dialogue systems with human-like dialogue capabilities.

This study has several limitations. First, the LLMs evaluated are limited to seven types, and it is unclear whether similar trends would be obtained when more diverse models and architectures are included. Second, Pearson correlation has several assumptions, and it is possible that the current experimental setting may not fully satisfy these assumptions. Therefore, it will be necessary to analyze the correlations in more detail by using other indicators (e.g., Spearman correlation) in the future. Third, the findings regarding the correlation between second- or higher-order ToM and dialogue ability are limited. The dialogue tasks utilized in this study are not considered to include tasks where second- or higher-order belief estimation abilities directly contribute to task achievement, and this constraint in task design is likely one of the reasons for the weak correlations observed in higher-order ToM. In the future, we plan to introduce dialogue tasks that require second- or higher-order belief

estimation. Additionally, in the Hi-ToM used in this study, the accuracy rate of tasks by humans has not been measured, and human performance on higher-order ToM reasoning tasks remains unclear. In the future, by measuring the accuracy rate of humans, we will be able to more accurately position the relationship between ToM for second- or higher-order beliefs and dialogue ability. Fourth, the dialogue tasks used in this study are limited to the five game-style tasks included in Clembench and the MutualFriends task, and it remains unclear whether the findings would generalize to tasks covering broader domains (such as MultiWOZ) or to non-task-oriented dialogues (such as casual conversation). Finally, it will also be necessary to determine how dialogue models can be systematically improved using the empirical insights gained from this study.

Acknowledgments

This work was supported by the "R&D Hub Aimed at Ensuring Transparency and Reliability of Generative AI Models" project of the Ministry of Education, Culture, Sports, Science and Technology.

References

Cristian-Paul Bara, Sky CH-Wang, and Joyce Chai. 2021. MindCraft: Theory of mind modeling for situated dialogue in collaborative tasks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1112–1125.

Simon Baron-Cohen, Alan M. Leslie, and Uta Frith. 1985. Does the autistic child have a "theory of mind" ? *Cognition*, 21(1):37–46.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026.

Kranti Chalamalasetti, Jana Götze, Sherzod Hakimov, Brielen Madureira, Philipp Sadler, and David Schlangen. 2023. clembench: Using game play to evaluate chat-optimized language models as conversational agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11174–11219.

Chunkit Chan, Cheng Jiayang, Yauwai Yim, Zheye Deng, Wei Fan, Haoran Li, Xin Liu, Hongming Zhang, Weiqi Wang, and Yangqiu Song. 2024. NegotiationToM: A benchmark for stress-testing machine theory of mind on negotiation surrounding. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4211–4241.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. 2024. ToMBench: Benchmarking theory of mind in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15959–15983.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press.

Uta Frith. 1994. Autism and theory of mind in everyday life. *Social Development*, 3(2):108–124.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. 2023. Understanding social reasoning in language models with language models. *In Proceedings of the 37th International Conference on Neural Information Processing Systems*, 36:13518–13529.

He He, Anusha Balakrishnan, Mihail Eric, and Percy Liang. 2017. Learning symmetric collaborative dialogue agents with dynamic knowledge graph embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1766–1776.

Yoshinobu Kano, Yuto Sahashi, Neo Watanabe, Kaito Kagaminuma, Claus Aranha, Daisuke Katagami, Kei Harada, Michimasa Inaba, Takeshi Ito, Hirotaka Osawa, Takashi Otsuki, and Fujio Toriumi. 2024. AIWolfDial 2024: Summary of natural language division of 6th international AIWolf contest. In *Proceedings of the 2nd International AIWolfDial Workshop*, pages 1–12.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023. FANToM: A benchmark for stress-testing machine theory of mind in interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14397–14413.

Jin-Hwa Kim, Nikita Kitaev, Xinlei Chen, Marcus Rohrbach, Byoung-Tak Zhang, Yuandong Tian, Dhruv Batra, and Devi Parikh. 2019. CoDraw: Collaborative drawing as a testbed for grounded goal-driven communication. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6495–6513.

Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*, 4:169.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint*

Conference on Natural Language Processing, pages 5872–5877.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132.

Xiaomeng Ma, Lingyu Gao, and Qihui Xu. ToMChallenges: A principle-guided dataset and diverse evaluation tasks for exploring theory of mind. In *Proceedings of the 27th Conference on Computational Natural Language Learning*, pages 15–26.

Michael McTear. 2022. *Conversational AI: Dialogue systems, conversational agents, and chatbots*. Springer Nature.

Shikib Mehri and Maxine Eskenazi. 2020. Unsupervised evaluation of interactive dialog with DialoGPT. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 225–235.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

François Quesque and Yves Rossetti. 2020. What do theory-of-mind tasks actually measure? theory and practice. *Perspectives on Psychological Science*, 15(2):384–396.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8689–8696.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2024. Clever hans or neural theory of mind? stress testing social reasoning in large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2257–2273.

Kazutoshi Shinoda, Nobukatsu Hojo, Kyosuke Nishida, Saki Mizuno, Keita Suzuki, Ryo Masumura, Hiroaki Sugiyama, and Kuniko Saito. 2025. ToMATO: Verbalizing the mental states of role-playing LLMs for benchmarking theory of mind. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1520–1528.

Takuma Udagawa and Akiko Aizawa. 2019. A natural language corpus of common grounding under continuous and partially-observable context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7120–7127.

Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649.

Heinz Wimmer and Josef Perner. 1983. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128.

Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. 2023. Hi-ToM: A benchmark for evaluating higher-order theory of mind reasoning in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10691–10706.

Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. 2024. OpenToM: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8593–8623.

Zihao Yi, Jiarui Ouyang, Yuwen Liu, Tianhao Liao, Zhe Xu, and Ying Shen. 2024. A survey on recent advances in LLM-based multi-turn dialogue systems. *arXiv preprint arXiv:2402.18013*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675*.

A Example Problems in ToM benchmarks

Table 9 presents example problems of the ToM benchmarks used in the study.

BenchMark	Context	QA
ToMBench (Chen et al., 2024)	Li Lei and Han Meimei are wandering in the living room, they see the cabinet, box and handbag, they find a sweatshirt in the cabinet, Han Meimei leaves the living room, Li Lei moves the sweatshirt to the handbag.	Question: Where is the sweatshirt now? (A) Briefcase (B) Box (C) Cabinet **(D) Handbag** Question: After Han Meimei returns to the living room, where does Li Lei think Han Meimei looks for the sweatshirt? (A) Box (B) Wardrobe (C) Handbag **(D) Cabinet**
FANTOM (Kim et al., 2023)	Kailey: Hey guys, I'll go grab a coffee. Sally: See you, Kailey! Hey Linda, did you get a dog? Linda: Yeah, I got a golden retriever. She's so adorable. ... Kailey: I'm back, what are you guys discussing now? Sally: Linda was just telling us that her dog can do special moves! Linda: Yeah, she can stand on her feet and do a dance move to music!	BeliefQA: What breed would Kailey think Linda's dog is? (A) Kailey believes Linda has a golden retriever. **(B) Kailey does not know the breed.** AnswerabilityQA: Who knows the correct answer to "What breed would Kailey think Linda's dog is"? **Linda, David, Sally** InfoAccessibilityQA: Who knows about "Linda has a golden retriever"? **Linda, David, Sally**
HI TOM (Wu et al., 2023)	William, Jack, Charlotte, Noah and Hannah entered the hall. Noah saw a monkey. The carrot is in the red_basket William exited the hall. ... Jack exited the hall. Charlotte exited the hall. Noah moved the carrot to the green_envelope. Noah exited the hall. Hannah moved the carrot to the red_basket. Hannah exited the hall. William, Jack, Charlotte, Noah and Hannah entered the waiting_room. Charlotte publicly claimed that carrot is in the green_envelope. Hannah privately told Charlotte that the carrot is in the blue_container.	Question-order0: Where is the carrot really? (A) green_envelope, **(B) red_basket**, ⋯ Question-order1: Where does William really think the carrot is? **(A) green_envelope**, (B) red_basket, ⋯ Question-order2: Where does Hannah think William thinks the carrot is? (A) green_envelope, **(B) red_basket**, ⋯ Question-order3: Where does Jack think Hannah thinks William thinks the carrot is? (A) green_envelope, **(B) red_basket**, ⋯ Question-order4: Where does Charlotte think Jack thinks Hannah thinks William thinks the carrot is? (A) green_envelope, **(B) red_basket**, ⋯

Table 9: Examples from the three ToM benchmarks addressed in this study. FANToM has three subtasks: BeliefQA, which directly estimates beliefs; AnswerabilityQA, which asks about the answerability of the questions; and InfoAccessibilityQA, which asks about people who know the information. In Hi-ToM, questions corresponding to 0th- to 4th-order ToM inference are set from Question-order 0 to 4. Bold portions in the QA items indicate the correct answers for each question.

B ToM Task Prompts

This section presents the prompts used by LLMs to solve ToMBench and FANToM. The prompt for solving ToMBench is as follows. `{context}` contains the context that serves as the basis for inference, `{question}` contains questions about characters' mental states, and `{a}`, `{b}`, `{c}`, and `{d}` are the answer choices.

```
Please read the passage and the question I will ask. Choose the correct answer from options A, B, C, and D.
{context}
{question}
A: {a}
B: {b}
C: {c}
D: {d}
Please answer with the letter of the option that you think is correct and do not output anything other than a single letter.
```

The following are the prompts used to solve FANToM, which is used for BeliefQ, InfoQ, and AnsQ. `{context}` contains the dialogue text that serves as the basis for inference, and `{BeliefQ}`, `{InfoQ}`, and `{AnsQ}` contain question texts defined for each task by the dataset. Additionally, `{factQ}` and `{factA}` contain the facts asked in BeliefQ, and `{candidates}` lists the names of the characters.

```
{context}
Question: {BeliefQ}
{ans_a}
{ans_b}
Please choose either a or b as the correct answer. Output only a or b.
```

```
{context}
Information: {factQ} {factA}
Question: {InfoQ}
Characters: {candidates}
Choose the characters who correctly answer the question from the list above.
Separate names with commas.
Answer:
```

```
{context}
Target: {factQ}
Question: {AnsQ}
Characters: {candidates}
Choose the characters who correctly answer the question from the list above.
Separate names with commas.
Answer:
```

C Dialogue Task Prompt

The prompts used in the MutualFriends task are as follows. Among these, `{subject}` and `{friends}` contain a list of friends given to the player, and `{history}` contains the dialogue history.

```
You are a smart cooperative agent named Alice.
You have many friends with different attributes (Alice's knowledge base).
You are now discussing this with Bob. He also has a list of friends.
You will talk to Bob for a maximum of 20 turns to find a mutual friend as quickly as possible.
You can ask him questions or provide information about your friends.
In addition, you should try to mention as few attributes and friends as possible.
{subject}
{friends}
Generate your next utterance based on the following dialogue history. If there is no dialogue history, generate the first
utterance. Output only your next utterance.
{history}
Alice:
```