Bridging the Gap between Theory of Mind and Action in LLMs

Anonymous ACL submission

Abstract

While Large Language Models (LLMs) demonstrate impressive capabilities in Theory of Mind (ToM) tasks, translating this social understanding into strategic action remains a critical challenge. In this paper, we study the ToM Knowing-Doing Gap, where models infer the mental states of others yet fail to select actions consistent with those inferences. To evaluate the discrepancy, we introduce the Theory of Mind-to-Action (ToMA) benchmark, a unified framework for analyzing the link between social reasoning and decision making in diverse cooperative and competitive scenarios. Our empirical analysis shows that correct ToM information does not always translate into strategic behavior, revealing failure modes where reasoning is decoupled from action. We further propose ToM-guided strategies that make ToM information explicit during action selection. Across models and scenarios, these strategies help mitigate the gap, suggesting that social intelligence in LLMs requires not only inferring others' mental states but also coupling those inferences with action.

1 Introduction

Large Language Models (LLMs) have recently achieved striking advances in complex reasoning and knowledge-intensive tasks (OpenAI, 2023; Chowdhery et al., 2022; Team, 2023; DeepSeek-AI et al., 2025). In parallel, a growing body of work has examined whether LLMs exhibit Theory of Mind (ToM)-like abilities. ToM refers to the ability to reason about others' mental states, including their beliefs, goals, and knowledge states (Premack and Woodruff, 1978; Baron-Cohen et al., 1985). Recent studies suggest that LLMs can often perform well on ToM tasks (Strachan et al., 2024; Kosinski, 2024; Chen et al., 2024; Marchetti et al., 2023). However, success on mental-state inference does not necessarily imply that these inferences are used to guide the model's own strategic actions.

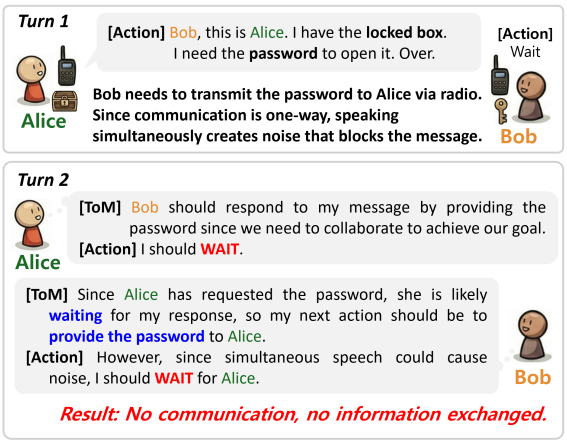

Figure 1: Example scenario of the Speak-or-Wait Game. At Turn 2, Alice successfully translates her ToM prediction that Bob will speak into the action of waiting for Bob. However, Bob fails to translate his ToM prediction that Alice will wait into the corresponding action of speaking to Alice, creating a ToM knowing–doing gap and resulting in failed information exchange.

Recent studies further reveal a critical failure mode known as the *Knowing-Doing Gap* (Schmied et al., 2025; Paglieri et al., 2024; Ruoss et al., 2024; Wong et al., 2025). Even when a model derives the correct inference, it often selects an action that contradicts that inference during decision making. This inconsistency undermines confidence in the reliability of LLM decisions and raises questions about their robustness, thereby limiting their potential as effective autonomous agents.

We specifically address this issue in the context of ToM, where inferred mental states must directly guide strategic choices. ToM is more than a passive diagnostic ability; it serves as a prerequisite for strategic action in social environments. It enables agents to coordinate toward shared goals in cooperative environments and to anticipate and counter opponents in competitive settings.

However, LLMs frequently fail to translate ac-

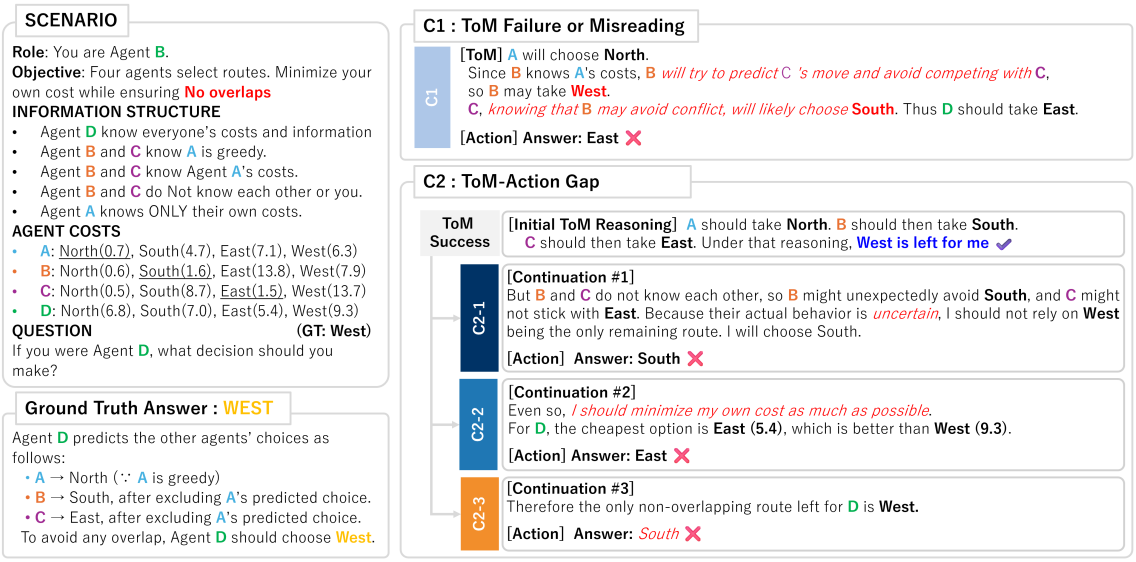

Figure 2: Representative examples of `C1` and `C2` failures in a PO scenario. `C1` captures task-understanding or ToM-inference failures, whereas `C2` captures ToM-to-action failures, including uncertainty-driven deviation (`C2-1`), prior or heuristic override (`C2-2`), and execution inconsistency (`C2-3`).

curate social inferences (*Knowing*) into appropriate behavioral choices (*Doing*). As illustrated in the one-way communication scenario in Figure 1, Bob remains silent despite correctly predicting that Alice awaits a response. This confirms that the *Knowing-Doing Gap* extends to social intelligence, revealing a critical disconnect between mental state reasoning and strategic action.

While prior research (Riemer et al., 2025) raised concerns about inconsistencies between ToM reasoning and action, it relies on toy cases and lacks rigorous validation. Furthermore, as highlighted by Riemer et al. (2025), mainstream ToM benchmarks focus on attributing mental states or predicting behavior of *other agents*, overlooking the agent's *own* choices. Consequently, the field lacks a systematic understanding of whether LLMs leverage social reasoning to guide their actions.

Contributions. This paper studies whether social reasoning in LLMs is translated into their own strategic actions. To this end, we introduce a benchmark for evaluating this connection, provide a mechanistic diagnosis of where it fails, and propose structured strategies that make ToM information more explicit during action selection. Our main contributions are as follows:

- **ToMA Benchmark:** We introduce the Theory of Mind-to-Action (ToMA) benchmark, a unified framework for systematically evaluating whether ToM-related inferences are translated into the model's own strategic actions across diverse cooperative and competitive scenarios.
- **Mechanistic Diagnosis:** We provide a diagnostic analysis of the ToM-to-action gap, showing that correct or near-correct ToM information does not always yield correct actions.
- **ToM-Conditioned Decision Strategies:** We propose `ToM-CoT` and `ToM-Guide` to mitigate the Knowing-Doing Gap by making ToM information more action-relevant.

2 Theory of Mind-to-Action (ToMA)

2.1 Overview

We introduce the **Theory of Mind to Action (ToMA)** benchmark, a unified framework evaluating whether LLMs translate social reasoning into strategic action. Unlike previous benchmarks treating ToM as a standalone classification task, ToMA ensures that solving the task remains strictly contingent on translating ToM inferences into action, thereby directly evaluating whether the model bridges social reasoning and behavior. The benchmark comprises four scenarios categorized into cooperative and competitive settings, as well as single-turn and multi-turn interactions.

2.2 Design Philosophy & Distinctions

While existing ToM benchmarks are limited to assessing the *presence* of social reasoning capabilities, ToMA shifts the focus to the *translation* of

this reasoning into behavior. We distinguish our approach through three key design principles:

- **Action-oriented strategic evaluation:** We shift the paradigm from passive mental state attribution (e.g., "What does Alice think?") to active strategic execution (e.g., "Given what Alice thinks, what will you do?"). This approach directly tests the alignment between social understanding and actual strategic behavior.
- **Deterministic causal dependency:** To ensure a rigorous link between reasoning and behavior, we design scenarios where correct ToM inference acts as the necessary and sufficient condition for the unique optimal action under strict conditions. This ensures that success reflects genuine alignment rather than coincidental task completion.
- **Complex multi-agent modeling:** ToMA simulates real-world social dynamics involving information asymmetry, strategic utility, and hidden agendas. Agents must synthesize divergent perspectives from partial observability to navigate complex interactions, such as coordinating with allies or deceiving adversaries.

2.3 Scenario Descriptions

ToMA consists of four distinct games, with a full scenario overview provided in Appendix A.

- **Partially Observable Game (PO):** This scenario centers on information asymmetry where selecting unique destinations others avoid requires the agent to link ToM reasoning directly to behavior. Identifying the optimal action remains impossible unless the agent translates perspective modeling into execution (Appendix A.1).
- **Minority Game (MG):** This game involves preference-based reasoning where identifying minority options depends strictly on translating social preference analysis into action. Success ensures the agent converts its understanding of others' likes into a final decision (Appendix A.2).
- **Speak or Wait (SW):** This multi-turn game requires the agent to lead allies to the target location via precise strategic communication while keeping adversaries strictly unaware of the final objective. Achieving the goal depends on applying ToM to avoid communication overlap with teammates and coordinate optimal timing while evading enemy detection (Appendix A.3).
- **Ambiguity Game (AG):** This multi-turn game focuses on strategic execution in mixed-motive settings. The agent must evaluate the mental states of both allies and adversaries to provide hints that selectively guide teammates while misleading enemies (Appendix A.4).

2.4 The ToM Knowing-Doing Gap

We identify a distinct manifestation of the *Knowing-Doing Gap* in social reasoning: the *ToM Knowing-Doing Gap*. Even when models can attribute mental states to others, they often fail to translate these inferences into strategically coherent actions.

Figure 2 illustrates representative failure patterns in ToMA scenarios. We categorize incorrect responses into two broad types. `C1` denotes failures in task understanding or ToM inference, including scenario misinterpretation, constraint miscalculation, or incorrect mental-state inference. `C2` denotes action-selection failures despite apparently successful ToM inference: the model identifies the relevant mental-state information, but this information does not reliably determine the final action.

We further divide `C2` failures into three subtypes:

- **`C2-1`: Uncertainty-driven deviation.** The model infers the ToM information but treats it as uncertain and chooses an alternative action. For example, it may infer another agent's likely behavior, yet decide that the agent might act differently or that the inference may be unreliable.
- **`C2-2`: Prior or heuristic override.** The model selects an action based on an intrinsic prior or heuristic rather than the ToM-consistent choice. In PO, this appears as greedy cost minimization, where the model chooses a lower-cost option despite violating the ToM-implied constraint. In MG, it appears as lexical or semantic bias, such as avoiding a crowded-sounding option like *Open Hall* despite its ToM-implied correctness.
- **`C2-3`: Execution inconsistency.** The model derives the ToM-consistent action during reasoning but outputs a different final answer.

We provide an empirical analysis of these error categories in Section 4.1, with additional analysis and experiments on `C2-2` in Appendix B.2.

Takeaway. Even when a model correctly infers others' mental states, such reasoning does not always translate into action. The ToM-to-Action Gap arises when the model treats its inference as uncertain, follows intrinsic priors such as cost minimization, or derives the correct action in reasoning but outputs a different final answer.

2.5 Evaluation Metrics

To quantify the gap between reasoning and action, we define three primary metrics based on the model's performance over N scenarios:

- **ToM Accuracy** (Acc_{ToM})**:** This metric evaluates the model's social reasoning capacity (*Knowing*). It measures the frequency with which the model correctly infers ground-truth mental states:

$$Acc_{ToM} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(\hat{M}_i = M_i) \quad (1)$$

  where $\hat{M}_i$ is the inferred mental state and M_i is the ground truth for the i-th scenario.

- **Action Consistency** (C_{act})**:** This metric assesses logical alignment between ToM reasoning and execution (*Doing*). It measures the frequency with which the model's chosen action A_i conforms to the valid action set $\mathcal{A}_{\hat{M}_i}$ derived from ToM inference. For example, if the model predicts a teammate will `WAIT` in SW, the set $\mathcal{A}_{\hat{M}_i}$ contains only `SPEAK` to ensure coordination. Selecting `SPEAK` as A_i increases the C_{act} value. The formula for C_{act} follows:

$$C_{act} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(A_i \in \mathcal{A}_{\hat{M}_i}) \quad (2)$$

- **Task Success Rate** (Acc_{task})**:** This metric represents the overall problem-solving performance. In scenarios like PO and MG, success is achieved only when the model correctly executes the optimal action A_i^* based on accurate ToM:

$$Acc_{task} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(A_i = A_i^*) \quad (3)$$

3 Enhancing ToM-to-Action Alignment

3.1 ToM-CoT

To strengthen the logical link between inference and action, we introduce `ToM-CoT`. Unlike standard CoT, `ToM-CoT` mandates a structured, two-stage sequential reasoning process:

$$P(A \mid X) \approx P(\hat{M} \mid X) \cdot P(A \mid \hat{M}, X) \quad (4)$$

where X represents the context, $\hat{M}$ denotes the explicitly articulated mental states, and A is the resulting action. As illustrated in Figure 3, this process structures the model's internal monologue into two distinct steps:

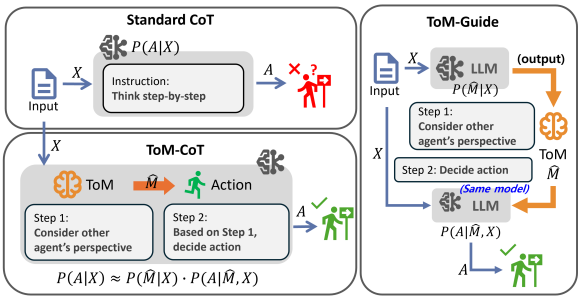

Figure 3: ToM driven architectures designed to improve consistency between reasoning and action

1. **ToM Reasoning:** The model infers beliefs, desires, and information constraints of all agents, generating a structured social inference $\hat{M}$.
2. **ToM-Guided Action:** The model derives the optimal action A by conditioning its decision-making process on the previously reasoned $\hat{M}$.

By eliciting an explicit path from "knowing others" to "deciding for oneself", `ToM-CoT` encourages social insights to serve as the basis for behavioral decisions. This structured reasoning reduces the likelihood that the model ignores its own ToM inference during final action selection.

3.2 ToM-Guide

While `ToM-CoT` internalizes the reasoning path within a single prompt, `ToM-Guide` externalizes the transition from reasoning to action through a two-step prompting pipeline. Both phases use the same underlying LLM rather than separate models; `ToM-Guide` differs from `ToM-CoT` by externalizing the inferred ToM state and re-injecting it in a second call. This approach encourages the model's behavioral policy to be explicitly conditioned on its previous social inferences. In the first phase, the model identifies the mental states $\hat{M}$ of the involved agents. In the second phase, the model receives both the original context X and its own inferred mental states $\hat{M}$ as augmented inputs to determine the final action:

$$A_{final} = \arg\max_{a \in \mathcal{A}} P(a \mid X, \hat{M}) \quad (5)$$

As illustrated in Figure 3, `ToM-Guide` makes the inferred ToM state an explicit premise for decision-making. By re-feeding $\hat{M}$ into the decision-making stage, `ToM-Guide` mitigates *the ToM Knowing-Doing Gap* by making final action selection explicitly conditioned on the previously articulated social reasoning.

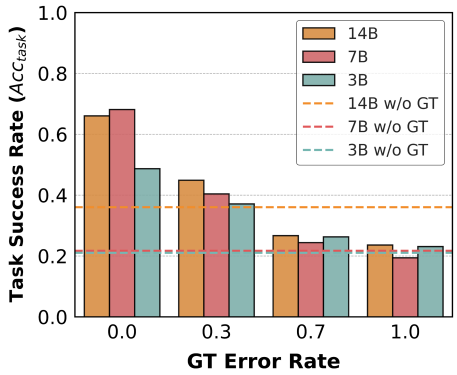

Figure 4: Scaling trends in average task success rates for the Qwen 2.5 series in PO and MG with ground truth ToM injection. The dotted line indicates the task success rate without GT injection.

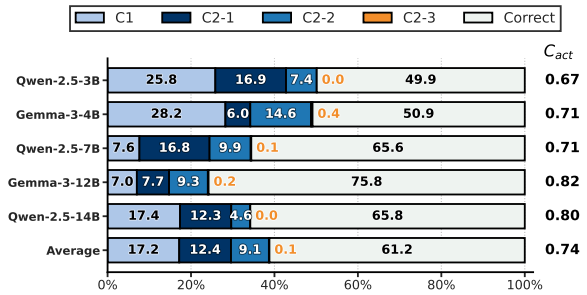

Figure 5: Breakdown of error categories in GT-injection experiments on PO and MG. Larger models show higher correct-action proportions and action consistency, suggesting better translation of ToM into action.

4 Experiments

In this section, we analyze the *ToM Knowing-Doing Gap* by examining the discrepancy between ToM inference and execution. We then evaluate how effectively `ToM-CoT` and `ToM-Guide` bridge this gap across strategic environments. Our experiments address the following research questions:

RQ1: The ToM Knowing-Doing Gap. To what extent does a discrepancy exist between a model's ToM inference and its actual behavioral execution?

RQ2: Strategy Effectiveness. How effectively do our proposed strategies mitigate the identified gap to improve strategic performance?

RQ3: Impact of ToM Complexity. How does the effectiveness of our proposed strategies vary across different levels of recursive social reasoning?

Models. We evaluate our methods across four distinct strategic scenarios within our proposed ToMA benchmark. We test various state-of-the-art LLMs to ensure generalizability, including the Qwen-2.5 series (Qwen et al., 2025), Gemma-3 series (Team et al., 2025), GPT-4o-mini(OpenAI, 2023), GPT-5.1 (OpenAI, 2025a), Gemini-2.0-flash (Team, 2023), and Gemini-2.5-flash-lite (Comanici et al., 2025). For these models, we operate them in instant mode to measure their direct response capabilities. Additionally, we include o4-mini (OpenAI, 2025b), which we execute in thinking mode to assess the impact of native long reasoning on strategic performance. We run all models five times and average their performance.

Baselines. To benchmark the performance of `ToM-CoT` and `ToM-Guide`, we compare them against three primary baselines: **Zero-shot**, where the model generates a final action A directly from the context X; **CoT**, which employs standard step-by-step reasoning without explicit ToM mandates; and **Self-consistency (SC)** (Wang et al., 2022), for which we perform 10 independent iterations and determine the final output through majority voting. For implementation, `ToM-CoT` enforces a two-stage internal monologue (ToM Reasoning followed by Action) in a single turn, while `ToM-Guide` utilizes a two-step pipeline that re-feeds the model's own inference as explicit context for the final decision.

4.1 Quantifying the ToM-Action Gap (RQ1)

We first quantify the ToM-to-action gap through a ground-truth injection (GT-injection) experiment. Instead of relying solely on the model's own ToM inference, we inject ToM information into the input prompt and examine whether the model uses this information to select the corresponding action. This setup allows us to separate failures in obtaining ToM information from failures in translating available ToM information into action.

Influence of ToM Accuracy. In the GT-injection experiment, we vary the GT error rate, which denotes the probability that the injected ToM information is replaced with an incorrect ToM signal. Figure 4 shows task success rates under different GT error rates. If ToM information were seamlessly translated into action, error-free GT injection would yield near-perfect task success. However, task success remains below 1.0 even at zero GT error, indicating the existence of the gap.

Error-type analysis. Figure 5 shows the breakdown of error categories in the GT-injection experiments on PO and MG. As model size increases, the correct-action proportion generally increases

Table 1: Comparison of task success rates (Acc_{task}) across LLM models and reasoning strategies in Partially Observable Games (PO). **Bold** indicates the best performance and underlining denotes the second best.

Type	Model	Acc_{ToM}	Baselines (Acc_{task})			Ours (Acc_{task})	
			Zero-shot	SC	CoT	ToM-Guide	ToM-CoT
Open	Qwen-2.5-3B-Instruction	0.327	0.283	0.284 (+0.00)	0.237 (-0.05)	0.292(+0.01)	**0.301** (+0.02)
	Qwen-2.5-7B-Instruction	0.378	0.258	0.224 (-0.03)	0.253 (-0.01)	0.279(+0.02)	**0.365** (+0.11)
	Qwen-2.5-14B-Instruction	0.442	0.211	0.213 (+0.00)	0.367 (+0.16)	0.404(+0.19)	**0.495** (+0.28)
	Qwen-2.5-32B-Instruction	0.487	0.284	0.289 (+0.01)	0.447 (+0.16)	0.462(+0.18)	**0.512** (+0.23)
	gemma-3-4b-it	0.323	0.255	0.227 (-0.03)	0.255 (+0.00)	**0.338** (+0.08)	0.300(+0.05)
	gemma-3-12b-it	0.364	0.294	0.299 (+0.01)	0.353 (+0.06)	0.407(+0.11)	**0.432** (+0.14)
	gemma-3-27b-it	0.427	0.262	0.263 (+0.00)	0.417 (+0.16)	0.448(+0.19)	**0.475** (+0.21)
Closed	Gemini-2.0-flash	0.507	0.410	0.441 (+0.03)	0.541 (+0.13)	**0.626** (+0.22)	0.565(+0.16)
	Gemini-2.5-flash-lite	0.511	0.394	0.433 (+0.04)	0.330 (-0.06)	**0.629** (+0.24)	0.556(+0.16)
	GPT-4o-mini	0.453	0.186	0.189 (+0.00)	0.301 (+0.12)	0.420(+0.23)	**0.464** (+0.28)
	GPT-5.1	0.604	0.301	0.309 (+0.01)	0.312 (+0.01)	0.639(+0.34)	**0.709** (+0.41)
	o4-mini (think)	-	0.750	0.824	-	-	-

Table 2: Comparison of task success rates (Acc_{task}) across LLM models and reasoning strategies in Minority Games (MG). **Bold** indicates the best performance and underlining denotes the second best.

Type	Model	Acc_{ToM}	Baselines (Acc_{task})			Ours (Acc_{task})	
			Zero-shot	SC	CoT	ToM-Guide	ToM-CoT
Open	Qwen-2.5-3B-Instruction	0.459	0.216	0.235 (+0.02)	0.183 (-0.03)	**0.276** (+0.06)	0.258(+0.04)
	Qwen-2.5-7B-Instruction	0.519	0.152	0.152 (+0.00)	0.181 (+0.03)	0.318(+0.17)	**0.369** (+0.22)
	Qwen-2.5-14B-Instruction	0.631	0.212	0.229 (+0.02)	0.353 (+0.14)	0.420(+0.21)	**0.468** (+0.26)
	Qwen-2.5-32B-Instruction	0.672	0.226	0.222 (-0.00)	0.391 (+0.17)	**0.547** (+0.32)	0.494(+0.27)
	gemma-3-4b-it	0.470	0.180	0.195 (+0.02)	0.241 (+0.06)	0.300(+0.12)	**0.342** (+0.16)
	gemma-3-12b-it	0.541	0.211	0.211 (+0.00)	0.366 (+0.16)	**0.589** (+0.38)	0.420(+0.21)
	gemma-3-27b-it	0.528	0.207	0.218 (+0.01)	0.343 (+0.14)	**0.521** (+0.31)	0.476(+0.27)
Closed	Gemini-2.0-flash	0.619	0.440	0.474 (+0.03)	0.409 (-0.03)	**0.660** (+0.22)	0.578(+0.14)
	Gemini-2.5-flash-lite	0.628	0.278	0.198 (-0.08)	0.206 (-0.07)	0.347(+0.07)	**0.493** (+0.22)
	GPT-4o-mini	0.554	0.168	0.182 (+0.01)	0.296 (+0.13)	0.482(+0.31)	**0.591** (+0.42)
	GPT-5.1	0.683	0.260	0.285 (+0.03)	0.313 (+0.05)	**0.700** (+0.44)	0.373(+0.11)
	o4-mini (think)	-	0.554	0.674	-	-	-

and action consistency (C_{act}) also improves, suggesting that larger models better translate ToM information into action. However, a substantial gap remains even under GT injection. On average, failures caused by ToM-to-action errors (C2 = C2-1 + C2-2 + C2-3, 21.6%) are more frequent than failures in task understanding or ToM inference (C1, 17.2%). Among the C2 subtypes, uncertainty-driven deviation (C2-1, 12.4%) accounts for the largest share, followed by prior or heuristic override (C2-2, 9.1%). Execution inconsistency (C2-3) is rare, indicating that most ToM-to-action failures arise during action selection rather than from a final answer slip.

4.2 Strategic Performance Gains (RQ2)

As shown in Table 1 and Table 2, our proposed strategies significantly outperform baselines across the evaluated models. We observe that models with higher ToM accuracy (Acc_{ToM}) exhibit more pronounced gains under our strategies. This correlation suggests that our methods help convert available social reasoning capabilities into actionable strategies. For instance, in MG, GPT-5.1 achieves a performance boost of +44%p, suggesting that strong social understanding is most useful when explicitly connected to an action policy.

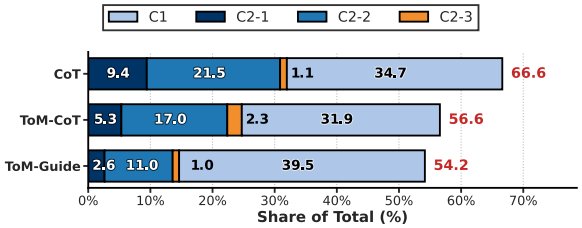

Figure 6: Breakdown of error categories across reasoning strategies on PO and MG. ToM-CoT and ToM-Guide reduce C2 failures compared with CoT, suggesting improved translation from ToM reasoning to action.

Figure 6 shows the distribution of error categories across reasoning strategies on PO and MG. Compared with CoT, both ToM-CoT and ToM-Guide reduce the overall error rate, from 66.6% to 56.6% and 54.2%, respectively. More

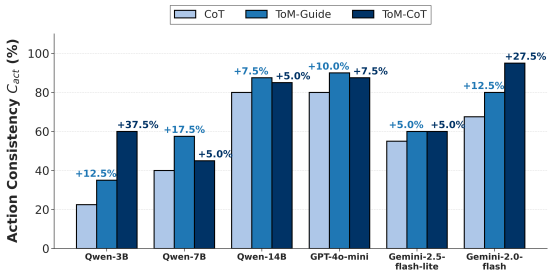

Figure 7: Comparison of average action consistency (C_{act}) across reasoning strategies in SW.

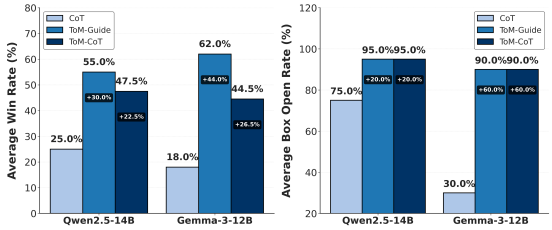

Figure 8: Average win and box open rates across reasoning strategies in the Speak or Wait (SW) game.

importantly, the reduction is concentrated in `C2` errors, which directly correspond to the ToM-to-Action Gap. While `C1` errors remain comparable or even increase slightly, `C2-1` and `C2-2` decrease substantially under our methods. This suggests that `ToM-CoT` and `ToM-Guide` are particularly effective at reducing failures where available ToM information is not translated into action.

In the SW game, our approach maintains higher Action Consistency (C_{act}) compared to the CoT baseline, which often fails to link its reasoning to behavioral output (Figure 7). Figure 8 further illustrates this impact through higher Box Open Rates and Win Rates. The increase in Box Open Rates indicates that our methods facilitate successful communication by aligning actions with social inferences about allies. Furthermore, the improved Win Rates imply that models successfully incorporate ToM reasoning about adversaries into their execution, allowing them to remain undetected while pursuing strategic goals.

We observe similar trends in the Ambiguity Game (AG), as presented in Figure 9. In this scenario, `ToM-Guide` and `ToM-CoT` consistently achieve higher win rates compared to the Zero-shot and CoT baselines. Notably, our strategies maintain a relatively low rate of `Leakage` and `Loss`, cases where the adversary exploits leaked information. The results indicate that our approach effectively translates dual social reasoning, involving cooperation with allies and concealment of intentions from opponents, into precise strategic actions.

4.3 Sensitivity to Reasoning Levels (RQ3)

We evaluate our strategies across various levels of recursive ToM reasoning, from non-social tasks to complex higher-order inferences. Figure 10 illustrates the performance breakdown by ToM level.

In Level 0 (L0) scenarios, where success depends on maximizing individual utility without considering others, we find that the Zero-shot baseline achieves highest accuracy. In this context, our strategies underperform because forcing social reasoning introduces unnecessary complexity to simple optimization problems.

However, as tasks demand recursive Theory of Mind (Level 1 and above), this trend reverses. While the CoT baseline outperforms Zero shot by introducing reasoning steps, it consistently falls short of our methods. Notably, as the ToM level increases towards Level 4, representing a fully observable and complex setting, the performance gap between our strategies and baselines widens. This suggests that while standard prompting methods struggle with higher-order social dynamics, our approach better mitigates the Knowing-Doing Gap. Thus, connecting ToM to action becomes increasingly important as strategic complexity grows, and our approach helps maintain this connection.

4.4 Preliminary Learning-based Adaptation

Beyond inference-time strategies, we examine whether the ToM-to-action gap can be mitigated through learning-based adaptation. As a proof-of-concept, we conduct a small-scale RL fine-tuning experiment in the SW setting. We train on data up to turn 2 and assign a reward of $+1$ when the model selects an action consistent with its ToM prediction, such as choosing SPEAK when it predicts the partner will WAIT, and -1 otherwise. After training, we evaluate the model in the full SW game; implementation details are provided in Appendix B.3.

As shown in Table 3, the RL-trained model improves over the baseline across all metrics, including Win Rate, Box Open rate, Acc_{ToM}, and C_{act}. These results suggest that learning-based adaptation can strengthen the coupling between ToM reasoning and action selection.

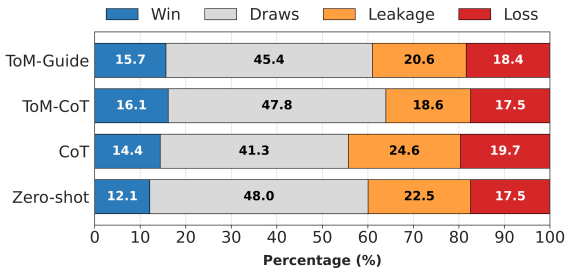

Figure 9: Comparison of execution results across reasoning strategies in the Ambiguity Game (AG). Refer to Figure 16 for specific results of each LLM model.

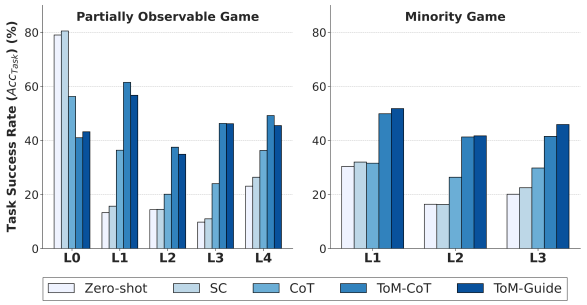

Figure 10: Comparison of task success rates (Acc_{task}) across strategies categorized by ToM level in Partially Observable and Minority Games.

Table 3: Performance comparison between the baseline and RL-trained method. Improvement denotes the absolute increase over the baseline in percentage points.

Method	Win Rate (%)	Box Open (%)	Acc_{ToM}	C_{act}
Baseline	11.67	43.33	20.37	27.02
RL-trained	21.67	51.67	24.03	31.41
Improvement	+10.00	+8.34	+3.66	+4.39

5 Related Work

ToM in LLMs and Benchmarks. Theory of Mind (ToM) refers to reasoning about others' mental states, such as beliefs, intentions, and knowledge states (Premack and Woodruff, 1978; Baron-Cohen et al., 1985). Early LLM evaluations adapted false-belief tests to assess mental-state attribution (Kosinski, 2024; Strachan et al., 2024), while later benchmarks such as **ToMi**, **OpenToM**, **FANToM**, **Hi-ToM**, and **ToMBench** evaluate belief tracking, information asymmetry, conversational ToM, and higher-order recursive reasoning (Le et al., 2019; Xu et al., 2024; Kim et al., 2023; He et al., 2023; Chen et al., 2024). **SimToM** further shows that modeling another agent's viewpoint can improve ToM performance in LLMs (Wilf et al., 2024). However, these works treat ToM as a passive diagnostic capability: they evaluate whether a model infers others' mental states or predicts another agent's behavior, but not whether those inferences guide the model's action selection.

Knowing-Doing and ToM-to-Action Gaps. Beyond social reasoning, recent studies show that LLMs may articulate correct rationales, beliefs, or values while acting inconsistently with them (Turpin et al., 2023; Pal et al., 2025; Shen et al., 2025; Schmied et al., 2025). Closer to our setting, **SimpleToM** studies the gap between explicit mental-state inference and downstream application (Gu et al., 2024), while **T4D** contrasts ToM-style inference questions with action-oriented responses (Zhou et al., 2023). Riemer et al. (2025) similarly argue that existing ToM benchmarks often assess the possession of ToM capabilities rather than their connection to behavior. These studies motivate the need to evaluate not only whether models can reason about others, but also whether such reasoning is behaviorally used.

Our work differs in its focus on the model's strategic action in controlled multi-agent settings. While open-ended multi-agent games such as *Werewolf* or *Diplomacy* implicitly require social reasoning (Xu et al., 2023; Park et al., 2023), their outcomes depend on confounding factors, making it difficult to isolate whether an action is driven by ToM reasoning. In contrast, ToMA designs scenarios where the optimal choice is tightly constrained by accurate social inference, enabling evaluation of whether ToM information is translated into action across cooperative and competitive scenarios.

6 Conclusion

In this paper, we study the *ToM Knowing-Doing Gap*, where LLMs infer others' mental states but fail to translate them into strategic actions. To evaluate this disconnect, we introduce the Theory of Mind-to-Action (ToMA) benchmark, which tests whether social reasoning guides action selection across cooperative and competitive scenarios. Our analysis shows that correct ToM information does not always yield coherent actions, revealing failure modes where social inference is decoupled from execution. We further show that ToM-CoT and ToM-Guide partially mitigate this gap by making ToM information explicit and action-relevant. Overall, our results suggest that social intelligence in LLMs requires not only inferring others' mental states, but also coupling those inferences with action.

Limitations

While our work offers significant insights into the transition from social reasoning to action, we acknowledge several limitations that future research should address.

First, although ToMA encompasses diverse social dynamics, the scenarios remain simplified abstractions of real-world interactions. In natural social settings, agents must process multimodal cues and navigate open-ended contexts where the optimal action is not always deterministic. Consequently, our benchmark might not fully capture the nuance of human-like social adaptability in unstructured environments.

Second, our proposed alignment strategies impose additional computational costs. `ToM-Guide`, in particular, requires a multi-step inference pipeline that generates explicit social reasoning before determining an action. This sequential process increases inference latency and token consumption compared to standard zero-shot prompting. Future work could explore more efficient methods, such as fine-tuning models on ToM-action pairs to internalize this alignment without the need for inference-time intervention.

Finally, our methods rely heavily on the underlying reasoning capability of the model. As observed in our experiments, alignment strategies yield minimal gains when the model lacks fundamental Theory of Mind accuracy. This dependency suggests that while we can bridge the gap between knowing and doing, we cannot compensate for a fundamental lack of social understanding.

Ethical Considerations

AI Assistance We utilized large language models, specifically GPT-5 (OpenAI, 2025a) and Gemini (Comanici et al., 2025), to assist in drafting and refining the clarity of this manuscript. Additionally, visual illustrations and figures were generated using Gemini Nano Banana Pro (Google, 2025). The authors reviewed and verified all content to ensure accuracy and assume full responsibility for the final text.

Dataset Generation and Content Safety The scenarios in the ToMA benchmark were constructed using a combination of human design and LLM assistance (e.g., GPT-4). We conducted a manual review of the dataset to ensure it does not contain harmful, violent, or offensive content. However, we acknowledge that LLM-generated text can occasionally produce unexpected outputs, and we encourage future research to maintain vigilance when employing similar data generation strategies.

Potential Biases Our benchmark aims to cover diverse social dynamics, including cooperation and competition. Nevertheless, the scenarios may reflect the implicit biases of the authors or the underlying models used during construction. We recognize that social norms vary across cultures, and the "optimal" actions defined in our benchmark may not universally apply to all cultural contexts.

Risks of Strategic Agents Finally, ToMA evaluates the ability of agents to take strategic actions, which includes deception in competitive settings (e.g., Minority Game). While our focus is on evaluating reasoning capabilities in controlled environments, we emphasize the importance of monitoring these behaviors to prevent potential misuse in real-world applications where honesty and safety are paramount.

References

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

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and 1 others. 2024. Tombench: Benchmarking theory of mind in large language models. *arXiv preprint arXiv:2402.15052*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, and 1 others. 2022. Palm: Scaling language modeling with pathways. *Preprint*, arXiv:2204.02311.

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*.

DeepSeek-AI, Daya Guo, Dejian Yang, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Google. 2025. Nano banana (image generation) | gemini api | google ai for developers. `https://ai.google.dev/gemini-api/docs/nanobanana`. Last updated 2025-12-18 UTC. Accessed 2026-01-06.

Yuling Gu, Oyvind Tafjord, Hyunwoo Kim, Jared Moore, Ronan Le Bras, Peter Clark, and Yejin Choi. 2024. Simpletom: Exposing the gap between explicit tom inference and implicit tom application in llms. *arXiv preprint arXiv:2410.13648*.

Yinghui He, Yufan Wu, 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. *Preprint*, arXiv:2310.16755.

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.

Michal Kosinski. 2024. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45):e2405460121.

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 (EMNLP-IJCNLP)*, pages 5872–5877.

Mengfan Li, Xuanhua Shi, and Yang Deng. 2025. Rectom: A benchmark for evaluating machine theory of mind in llm-based conversational recommender systems. *Preprint*, arXiv:2511.22275.

Antonella Marchetti, Cinzia Di Dio, Angelo Cangelosi, Federico Manzi, and Davide Massaro. 2023. Developing chatgpt's theory of mind. *Frontiers in Robotics and AI*, 10:1189525.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

OpenAI. 2025a. GPT-5.1 Instant and GPT-5.1 Thinking System Card Addendum. https://cdn.openai.com/pdf/4173ec8d-1229-47db-96de-06d87147e07e/5_1_system_card.pdf. Accessed 2026-01-05.

OpenAI. 2025b. Openai o3 and o4-mini system card. System card, OpenAI. Published April 16, 2025. Accessed 2026-01-05.

Davide Paglieri, Bartłomiej Cupiał, Samuel Coward, Ulyana Piterbarg, Maciej Wolczyk, Akbir Khan, Eduardo Pignatelli, Łukasz Kuciński, Lerrel Pinto, Rob Fergus, Jakob N. Foerster, Jack Parker-Holder, and Tim Rocktäschel. 2024. Balrog: Benchmarking agentic llm and vlm reasoning on games. *Preprint*, arXiv:2411.13543.

Arka Pal, Teo Kitanovski, Arthur Liang, Akilesh Potti, and Micah Goldblum. 2025. Incoherent beliefs & inconsistent actions in large language models. *Preprint*, arXiv:2511.13240.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Matthew Riemer, Zahra Ashktorab, Djallel Bouneffouf, Payel Das, Miao Liu, Justin D. Weisz, and Murray Campbell. 2025. Position: Theory of mind benchmarks are broken for large language models. *Preprint*, arXiv:2412.19726. ICML 2025.

Anian Ruoss, Fabio Pardo, Harris Chan, Bonnie Li, Volodymyr Mnih, and Tim Genewein. 2024. Lmact: A benchmark for in-context imitation learning with long multimodal demonstrations. *Preprint*, arXiv:2412.01441.

Thomas Schmied, Jörg Bornschein, Jordi Grau-Moya, Markus Wulfmeier, and Razvan Pascanu. 2025. Llms are greedy agents: Effects of rl fine-tuning on decision-making abilities. *Preprint*, arXiv:2504.16078.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Hua Shen, Nicholas Clark, and Tanu Mitra. 2025. Mind the value-action gap: Do LLMs act in alignment with their values? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3097–3118, Suzhou, China. Association for Computational Linguistics.

James W. A. Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessio Rufo, Stefano Panzeri, Federico Manzi, Michael S. A. Graziano, and Cristina Becchio. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7):1285–1295.

Gemini Team. 2023. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev,

and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Alex Wilf, Sihyun Lee, Paul Pu Liang, and Louis-Philippe Morency. 2024. Think twice: Perspective-taking improves large language models' theory-of-mind capabilities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8292–8308.

Annie Wong, Thomas Bäck, Aske Plaat, Niki van Stein, and Anna V. Kononova. 2025. Towards a deeper understanding of reasoning capabilities in large language models. *Preprint*, arXiv:2505.10543.

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. *arXiv preprint arXiv:2402.06044*.

Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. 2023. Exploring large language models for communication games: An empirical study on werewolf. *arXiv preprint arXiv:2309.04658*.

Pei Zhou, Aman Madaan, Srividya Pranavi Potharaju, Aditya Gupta, Kevin R McKee, Ari Holtzman, Jay Pujara, Xiang Ren, Swaroop Mishra, Aida Nematzadeh, and 1 others. 2023. How far are large language models from agents with theory-of-mind? *arXiv preprint arXiv:2310.03051*.

A Detailed Description of the ToMA Benchmark Scenarios

Figure 11 provides an overview of the ToMA Benchmark scenario. We plan to release the benchmark dataset upon publication. The license and terms of use will be specified in the release documentation.

A.1 Partially Observable Game (PO)

A.1.1 Environment

- **Agent Set:** A team consisting of four independent decision-makers, denoted as $\mathcal{A} = \{A, B, C, D\}$.
- **Task Set:** A set of four unique tasks $\mathcal{T} = \{T_1, T_2, T_3, T_4\}$ that must be performed.
- **Cost Matrix:** Defined as the cost $C_{i,j}$ incurred when agent i performs task j.
- **Information Structure:** Asymmetric access to cost information and meta-knowledge regarding others' knowledge states (Levels 0-4).

A.1.2 Dataset Description

The dataset consists of scenarios structured as described in Table 4. Each entry provides the context, cost data, and specific questions to test Theory of Mind reasoning. The dataset consists of 150 scenarios balanced across five reasoning levels.

A.1.3 Constraints

- **Exclusive Assignment:** Each task must be assigned to exactly one agent.
- **Overlap Penalty:** If two or more agents select the same task, the mission is immediately classified as a **"Mission Failure"**.
- **No Communication:** Agents must reason solely based on the provided information structure without any communication.

A.1.4 Objectives

- **Individual Optimization (Levels 0–3):** Minimize the agent's own cost while avoiding conflicts.
- **Social Optimization (Level 4):** Select the *Global Optimum* that minimizes the team's total cost: $\min \sum_{i \in \mathcal{A}} C_{i,\sigma(i)}$, where $\sigma(i)$ denotes the task assigned to agent i, and $C_{i,\sigma(i)}$ represents the cost incurred by agent i to perform that task.
- **Failure Prediction:** Detect if a collision among other agents is logically inevitable.

A.1.5 Reasoning and Strategy

Table 5 summarizes the ToM reasoning required per level.

A.2 Minority Game (MG)

A.2.1 Environment

- **Composition:** N agents ($N \in \{4, 6, 8\}$) and four locations ($L_1, \ldots, L_4$).
- **Information Structure:** Locations have unique *Features*. Agents have *Feature Preferences* (Likes/Dislikes) and *Social Relationships* (Follow/Avoid).

A.2.2 Dataset Description

The dataset defines scenarios where agents select locations based on preferences and social cues. The

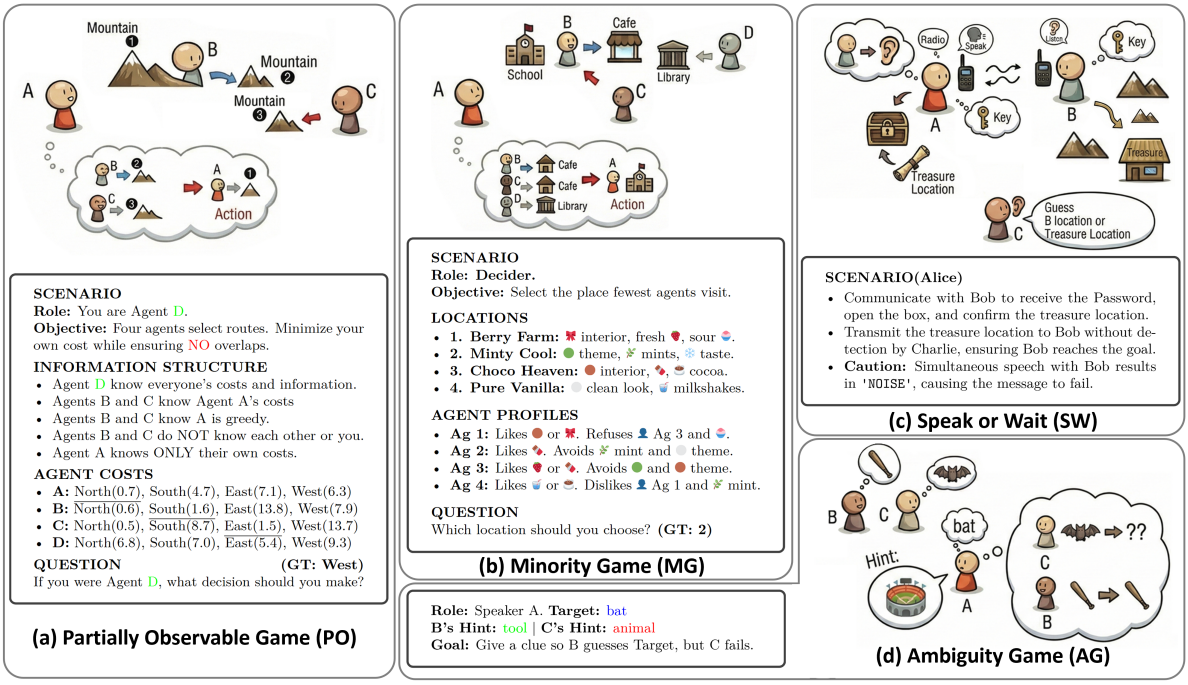

Figure 11: Overview of the four ToMA scenarios: Partially Observable Game, Minority Game, Speak-or-Wait, and Ambiguity Game.

Table 4: Dataset Structure for Partially Observable Game

Field	Description
`id`	Unique identifier for the scenario (e.g., L0_1).
`level`	ToM reasoning level required (0–4).
`context`	Textual description of the scenario, agent role, and objective.
`information_structure`	Description of cost visibility constraints.
`data`	JSON object containing cost values for each agent (A–D) per task.
`tom_questions`	List of questions testing belief prediction of other agents' actions.
`question`	The main decision-making query for the user/model.
`options`	List of possible actions (Task selection).
`answer`	Index of the optimal action (Ground Truth).

Table 5: ToM Requirements and Action Strategies per Level

Level	ToM Requirement	Action Strategy
Level 0	**Zero-order:** Focuses solely on own cost.	Select min C_{self}.
Level 1	**1st-order:** Predicts greedy moves of others.	Predict actions of known agents to adapt; ignore unknown agents.
Level 2	**2nd-order:** Recognizes knowledge chains.	Calculate downstream decisions and optimize remaining slot.
Level 3	**3rd-order:** Tracks "thoughts about thoughts".	Simulate recursive paths to find vacant tasks.
Level 4	**Cooperative:** Assumes mutual trust and full transparency.	Accept higher individual cost for team benefit (**Global Coordination**).

structure is detailed in Table 6. The dataset includes 90 scenarios divided into three levels of recursive reasoning.

A.2.3 Constraints

- **Information Restriction:** Agents only know social preferences of specific targets; no global population visibility.
- **Non-cooperative Environment:** Strict prohibition of communication; independent decision-making.

Table 6: Dataset Structure for Minority Game

Field	Description
id	Unique identifier for the scenario.
level	ToM reasoning level required.
context	Scenario setting (e.g., Cafe, Gym, Theme Park).
data	List of available locations and their specific feature attributes.
agents	Profiles of agents describing their likes (priorities), dislikes (deal-breakers), and social dependencies (Follow/Avoid).
tom_questions	Questions asking to predict the destination of specific agents.
question	The main query to identify the location with the minimum number of agents.
options	List of locations to choose from.
answer	Index of the location chosen by the minority (Global Minimum).

Table 7: Dataset Structure for Ambiguity Game

Field	Description
id	Unique identifier for the instance.
target	The target word (Ground Truth) that needs to be communicated by the Speaker.
b_hint	The private hint visible only to the Teammate (Agent B). The clue must align with this.
c_hint	The private hint visible only to the Opponent (Agent C). The clue must conflict with or be orthogonal to this.

A.2.4 Objectives

- **Minority Choice:** Identify and select the location with the **Minimum** number of agents.
- **Reasoning Goal:** Decipher dependency chains to predict the final destination of every agent.

A.2.5 Reasoning and Strategy

Agents perform stepwise ToM reasoning based on their assigned **Role**:

- **Anchor (Level 1):** Determines destination by comparing location *Features* only. (Prioritize positive, reject negative).
- **Intermediate (Level 2):** Predicts the **Anchor**'s destination, then decides to Follow or Avoid based on preferences.
- **Leaf (Level 3):** Predicts the **Intermediate**'s decision (which depends on the Anchor) and executes the final action based on this recursive logic.

A.3 Speak or Wait (SW)

A.3.1 Environment

- **Setting:** A 4×4 Grid World with named landmarks.
- **Agents:**
  - Agent A (Guide): Stationary, holds the locked box/map.
  - Agent B (Explorer): Mobile, holds the password.
  - Agent C (Adversary): Eavesdrops on transmissions.

A.3.2 Constraints

- **Half-duplex Channel:** Simultaneous SPEAK actions result in "NOISE" (collision).

A.3.3 Objectives

- **Primary Goal:** Securely exchange the password and location to guide Agent B to the target.
- **Optimization:** Minimize communication collisions and turn counts.

A.3.4 Reasoning and Strategy

Without prior protocol arrangement, agents must use Theory of Mind to resolve the "Speak or Wait" dilemma. Agents simulate the partner's intent to decide whether to transmit information or yield the channel, thereby establishing an implicit turn-taking order.

A.4 Ambiguity Game (AG)

A.4.1 Environment

- **Composition:** A Speaker (A), a Teammate (B), and an Opponent (C).
- **State:** Defined by $\mathcal{S} = (w, h_B, h_C)$, where w is the target word, h_B is B's hint, and h_C is C's hint.

A.4.2 Dataset Description

This dataset contains word association tasks where a single target word must be conveyed ambiguously. The fields are described in Table 7. The dataset contains 100 instances of word association tasks requiring targeted ambiguity.

A.4.3 Constraints

- **Information Asymmetry:** Speaker A sees full state. Listeners B and C only see their private hints and the public clue.
- **One-way Communication:** Only A broadcasts a single clue c. No feedback permitted.

A.4.4 Objectives

- **Knowledge Gap Exploitation:** Generate a clue c^* that maximizes Teammate decoding accuracy while minimizing Opponent accuracy.
- **Formal Condition:** $\text{prediction}(B \mid h_B, c) = w \wedge \text{prediction}(C \mid h_C, c) \neq w$.

A.4.5 Reasoning and Strategy

The Speaker must model the distinct perspectives of both listeners. By calculating the semantic distance between the potential clue and the divergent hints (h_B vs. h_C), the agent constructs an ambiguity that is resolved only when combined with the Teammate's private knowledge.

B Experiment Details

B.1 General Setting

We evaluate the performance of ToM-CoT and ToM-Guide on the ToMA benchmark using a diverse set of Large Language Models. We employ both open source and closed source models to ensure a comprehensive validation.

For the open source models, we utilize the Qwen2.5 series (Qwen et al., 2025) and the Gemma-3 series (Team et al., 2025). For the closed source models, we access GPT-4o-mini (OpenAI, 2023), GPT-5.1 (OpenAI, 2025a), o4-mini, gemini-2.5-flash-lite (Comanici et al., 2025), and gemini-2.0-flash (Team, 2023) via their official APIs.

We execute the experiments with a consistent set of hyperparameters to ensure fair comparisons. We set the temperature to 0.7 and the top-p value to 0.95. Regarding the maximum generation length, we adjust the maximum token parameter based on the specific sub-tasks within the benchmark. Specifically, we allocate 2048 tokens for PO and MG, 1024 tokens for SW, and 512 tokens for AG. Table 8 summarizes the hyperparameter configurations used in our experiments.

Parameter	Value
Temperature	0.7
Top-p	0.95
Max Tokens (PO, MG)	2048
Max Tokens (SW)	1024
Max Tokens (AG)	512

Table 8: Hyperparameter configurations used in our experiments.

B.2 Additional Error Analysis

B.2.1 Error Type Definitions

We categorize incorrect responses according to whether the error reflects a failure of ToM inference or a failure to translate ToM inference into action. `C1` captures failures in task understanding or ToM inference. These include cases where the model misreads the scenario, miscomputes constraints such as overlap or cost, or fails to infer the relevant mental state. In contrast, `C2` captures ToM-to-action failures: the model identifies the relevant ToM information, but this information does not reliably determine the final action.

We further divide `C2` into three subtypes. `C2-1` denotes uncertainty-driven deviation, where the model infers the relevant ToM information but treats it as unreliable and chooses an alternative action. `C2-2` denotes prior or heuristic override, where the model follows an intrinsic prior, such as cost minimization or lexical preference, rather than the ToM-consistent action. `C2-3` denotes execution inconsistency, where the model derives the ToM-consistent action during reasoning but outputs a different final answer.

B.2.2 Error Type Annotation Setup

We annotate incorrect responses by examining the model's reasoning trace and final answer. Each error is assigned to one of four categories: `C1`, `C2-1`, `C2-2`, or `C2-3`. We first use an LLM-based annotator to assign a preliminary label according to the criteria below, and then manually verify the assigned labels to ensure consistency. `C1` is used when the error originates from task misunderstanding or failed ToM inference. This includes cases where the model misreads what other agents know or do not know, constructs incorrect predictions of their actions, uses the information structure incor-

rectly, or provides reasoning too weak to establish a clear ToM-to-action chain.

For errors where the model shows some valid ToM or predicted-action reasoning, we assign one of the C2 subtypes. C2-1 captures safety or uncertainty override: the model derives a plausible ToM-based action, but deviates from it because it treats the inference as uncertain, risky, or not guaranteed. Typical indicators include phrases such as "to be safe," "safest," "might choose differently," or "cannot be certain." C2-2 captures prior-biased override: the model follows an intrinsic prior or heuristic despite its ToM reasoning. Common indicators include cost-based preferences such as "lowest cost," "cheapest," or "minimize my cost," as well as semantic priors about locations, such as assuming that a place is popular, crowded, quiet, or likely to be avoided.

Finally, C2-3 captures reasoning–answer inconsistency. We use this label when the reasoning trace supports a ToM-consistent action, but the final answer selects a different option, omits the supported action, or otherwise contradicts the stated conclusion. When multiple criteria appear in the same response, we assign the label based on the dominant source of the final error, prioritizing whether the final answer reflects failed ToM inference (C1), uncertainty override (C2-1), prior override (C2-2), or reasoning–answer mismatch (C2-3).

B.2.3 Intrinsic Priors(C2-2) as a Driver of ToM-to-Action Failures

Our error analysis suggests that intrinsic priors often dominate final action selection in C2 failures. In PO, this appears as greedy cost minimization. Even when ToM reasoning implies a unique valid action, models can drift toward a lower-cost option, sometimes one already selected by another agent, thereby violating the no-overlap constraint. This behavior corresponds to C2-2, where a heuristic preference overrides the ToM-consistent action.

We observe a complementary pattern in MG, where lexical or semantic biases influence action selection. For example, when an option such as *Focus Room* appears, models over-select it even when ToM reasoning implies that it should not be chosen. In our analysis, *Focus Room* is selected 77% of the time, corresponding to approximately $3.1\times$ over-selection. Conversely, crowded-sounding options such as *The Open Hall* can be avoided even when ToM predicts that they should be the correct minority choice. These patterns suggest that ToM-to-action failures are not limited to arithmetic or constraint-checking errors, but can also arise when language-mediated priors override socially grounded reasoning.

B.2.4 Controlled PO Cost-Flip Experiment

To further test whether intrinsic priors drive ToM-to-action failures, we construct a controlled PO cost-flip experiment with 50 matched instances. Figure 2 provides an illustrative example of this manipulation: ToM reasoning implies that agent D should select *West*. In the original condition, *West* is the unique ToM-consistent valid action, but its cost is relatively high, e.g., 9.3, and therefore it is not the minimum-cost option. Models may correctly identify that *West* is the valid remaining choice, yet still select a cheaper invalid option.

For each instance, we create a matched cost-flipped condition by changing the cost of the ToM-implied correct option. For example, in the Figure 2 scenario, we modify the cost of *West* for agent D so that it becomes the minimum-cost option, e.g., changing it from 9.3 to 5.3, while keeping the ToM structure, other agents' predicted actions, and the valid action unchanged. Thus, each pair differs only in whether the ToM-consistent action is aligned with the model's cost-minimization prior.

Using gemma-3-4b-it under GT injection, accuracy is 0.175 when the correct option is not the minimum-cost choice, but increases to 0.65 when the same ToM-implied option becomes the minimum-cost choice. This controlled flip indicates that models are more likely to follow ToM reasoning when it agrees with their intrinsic cost prior. When ToM reasoning conflicts with the prior, the model often selects a cheaper but invalid option, yielding C2-2 prior override or, in some cases, C2-3 execution inconsistency.

These results support the interpretation that a substantial portion of the ToM-to-Action Gap arises not because the model lacks ToM information, but because intrinsic priors can override that information during action selection.

B.3 RL Fine-tuning Details

We conduct a preliminary RL fine-tuning experiment to examine whether the ToM-to-action gap can be mitigated through learning-based adaptation. We use the Speak-or-Wait (SW) game, a multi-turn communication scenario where agents must coordinate with allies while avoiding information leakage to adversaries. Success in SW requires the model

to use ToM reasoning to decide whether to SPEAK or WAIT, avoiding communication overlap with teammates while preserving strategic concealment.

For training, we use a simplified SW setting that includes only the turn-2 decision. This isolates the core ToM-to-action transition: given the model's prediction of the partner's behavior, the model must select the complementary action. We fine-tune `gemma-3-4b-it` with PPO (Schulman et al., 2017) using a binary reward based on ToM-action consistency. Specifically, the model receives a reward of $+1$ if its action is consistent with its ToM prediction, and -1 otherwise:

$$r = \begin{cases} +1, & \text{if } A = A_{\hat{M}}, \\ -1, & \text{otherwise,} \end{cases}$$

where $\hat{M}$ denotes the model's inferred ToM state and $A_{\hat{M}}$ denotes the ToM-consistent action implied by that inference. For example, if the model predicts that the partner will WAIT, then choosing SPEAK receives $+1$; if it predicts that the partner will SPEAK, then choosing WAIT receives $+1$.

After training, we evaluate the RL-trained model on the full SW game rather than the simplified turn-2 setting. This tests whether learning ToM-action consistency at turn 2 transfers to complete multi-turn interaction. We compare the RL-trained model against the same base model without RL fine-tuning. For each condition, we run three evaluation batches with 20 game executions per batch, resulting in 60 full-game runs in total.

Table 3 presents Win Rate, Box Open rate, ToM accuracy (Acc_{ToM}), and action consistency (C_{act}). The RL-trained model improves over the baseline across all metrics, suggesting that training on local ToM-action consistency can strengthen the model's ability to connect social reasoning with strategic action selection in the full game. Since the reward directly optimizes only turn-2 consistency rather than full-game success, these results provide preliminary evidence that learning-based adaptation can improve ToM-to-action coupling beyond inference-time prompting.

B.4 Turn-2 Alignment Analysis in SW

We further analyze the alignment between ToM predictions and agent actions in the turn-2 decision of the Speak-or-Wait (SW) game. In this setting, a ToM-consistent action is determined by the predicted behavior of the partner: if the model predicts that the partner will WAIT, the aligned action is to SPEAK; if it predicts that the partner will SPEAK, the aligned action is to WAIT. We measure this alignment using action consistency (C_{act}).

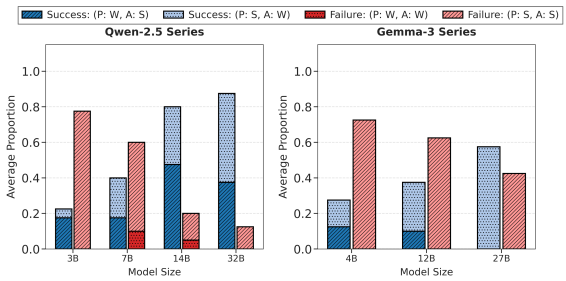

Figure 12: Alignment analysis between ToM predictions (P) and agent actions (A). Blue and red bars represent successful and failed alignment, respectively. W (Wait) and S (Speak) denote specific behaviors, such as an agent choosing to speak ($A : S$) when predicting the partner will wait ($P : W$).

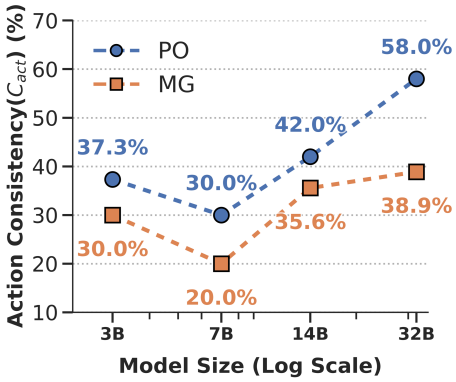

Figure 13: Action consistency of the Qwen2.5 series in Partially Observable Games and Minority Games.

Figure 12 compares turn-2 alignment across the Qwen-2.5 and Gemma-3 model families. The blue bars indicate cases where the model's action is consistent with its ToM prediction, while the red bars indicate misalignment. Across both model families, larger models generally show higher alignment, suggesting that model scale improves the ability to translate ToM predictions into action. However, misaligned cases remain even in larger models, indicating that ToM prediction and action selection are still not fully coupled.

B.5 C_{act} Scaling in PO and MG

Figures 14 and 13 compare action consistency (C_{act}) in PO and MG for the Gemma-3 and Qwen-2.5 series, respectively. In both model families, C_{act} generally increases with model size, suggesting that larger models better translate ToM reason-

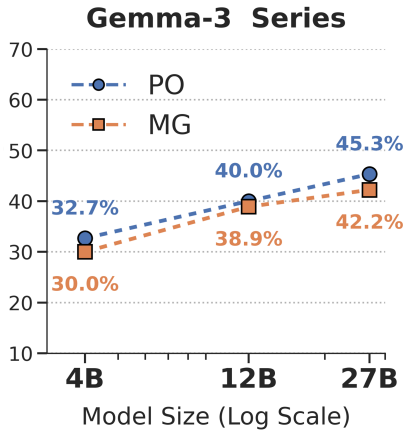

Figure 14: Action consistency of the Gemma 3 series in Partially Observable Games and Minority Games.

Table 9: Comparison of task success rates (Acc_{task}) on the RecToM *Recommender-prediction* scenario (Li et al., 2025), designed to evaluate ToM-to-action capabilities, and the IPD task by Riemer et al. (2025).

Task	Model	Baselines (Acc_{task})		Strategies (Acc_{task})	
		Zero-shot	CoT	ToM-CoT	ToM-Guide
RecToM	Gemma-3-12b-it	0.147	0.132 (-0.02)	**0.178** (+0.03)	0.165(+0.02)
	Qwen2.5-14B-Instruct	0.161	0.144 (-0.02)	**0.210** (+0.05)	0.189(+0.03)
	gemini-2.5-flash-lite	0.099	0.109 (+0.01)	**0.239** (+0.14)	0.187(+0.09)
IPD	Qwen2.5-3B-Instruct	0.000	0.355 (+0.36)	0.386(+0.39)	**0.507** (+0.51)

ing into action.

This trend supports our main observation that scale improves the translation from ToM inference to action selection. However, action consistency remains imperfect even for larger models, suggesting that scaling alone does not fully eliminate the ToM-to-Action Gap.

B.6 Generality Beyond ToMA

We examine whether ToM-CoT and ToM-Guide generalize beyond ToMA using two existing benchmarks, RecToM and IPD. Table 9 compares Zero-shot, CoT, ToM-CoT, and ToM-Guide across these settings.

Both methods generally outperform Zero-shot and CoT, suggesting that explicitly linking ToM reasoning to action selection can improve performance beyond our proposed benchmark. This supports the broader applicability of ToM-conditioned action selection, although the gains vary across tasks and models.

B.7 SW Results Across Teammate and Opponent Models

Figure 15 compares average Win Rate and Box Open rate in the SW game across different teammate and opponent models. The results show that ToM-CoT and ToM-Guide generally improve both metrics over the baselines, indicating that the proposed strategies are not tied to a single partner or adversary model.

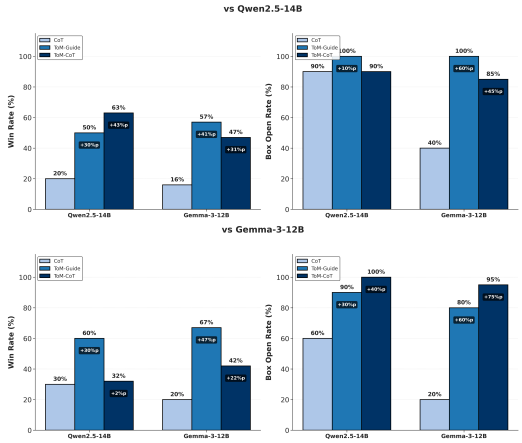

Figure 15: Average win rates and box open rates in the Speak or Wait game across various opponent and teammate models.

The gains in Box Open rate suggest that our methods improve cooperative coordination with teammates, while the gains in Win Rate indicate that they also help maintain strategic concealment against opponents. These results support the robustness of ToM-conditioned action selection in multi-agent interaction.

B.8 AG Results Across Teammate and Opponent Models

Figure 16 analyzes execution results in the Ambiguity Game (AG) across different teammate and opponent models. ToM-CoT and ToM-Guide generally show higher success rates than the baselines, suggesting that ToM-conditioned action selection remains effective in mixed-motive settings that require both teammate guidance and opponent concealment.

B.9 Level-wise Task Success in PO and MG

Table 10 presents level-wise task success rates on PO. Across most models, ToM-CoT and ToM-Guide achieve the best or second-best total performance, with especially clear gains in higher ToM levels. In contrast, Zero-shot and Self-consistency often perform strongly on L0, where success mainly depends on individual cost minimization rather than social reasoning, but their performance drops sharply in levels requiring recursive ToM. These results suggest that explicit ToM-conditioned reason-

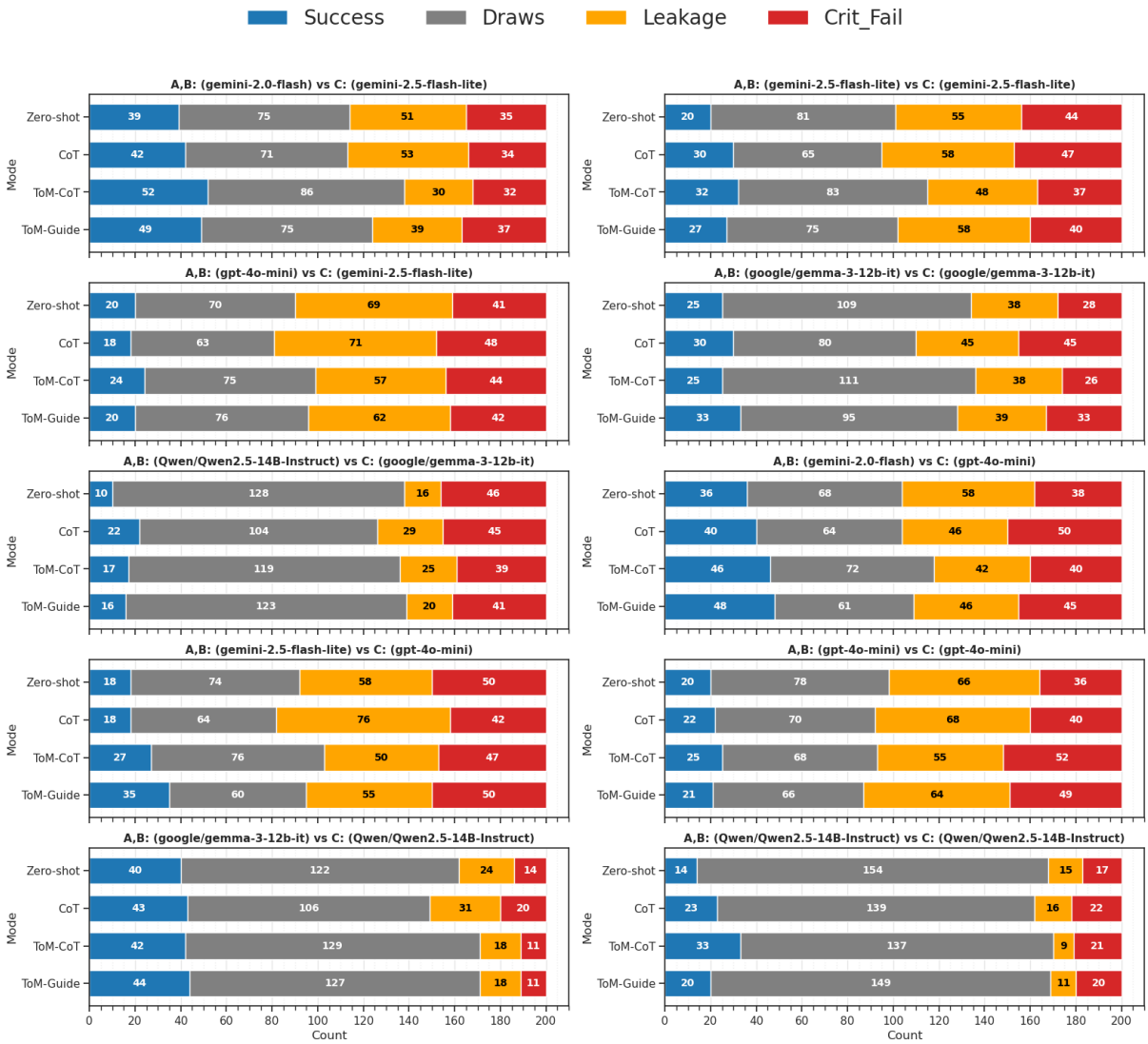

Figure 16: Analysis of execution results in the Ambiguity Game across various opponent and teammate models.

ing becomes increasingly important as PO requires more complex perspective tracking.

Table 11 shows level-wise task success rates on MG. Similar to PO, `ToM-CoT` and `ToM-Guide` generally outperform Zero-shot, Self-consistency, and CoT across L1–L3, with particularly strong gains for larger open models and closed models. The improvements are consistent across reasoning levels, indicating that our methods help models convert social preference reasoning into action choices rather than merely improving a single difficulty level. These results further support the role of ToM-conditioned action selection in preference-based multi-agent settings.

Table 10: Level-wise Task Success Rate (Acc_{task}) on Partially Observable Games. Orange shading indicates the best result, and light blue shading indicates the second best.

Family	Model	Strategy	Total	L0	L1	L2	L3	L4
Open Model								
Qwen	Qwen2.5-3B-Instruct	Zero-shot	0.283	0.607	0.114	0.287	0.120	0.287
		Zero-shot-SC	0.284	0.600	0.120	0.280	0.100	0.320
		CoT	0.237	0.547	0.154	0.153	0.114	0.220
		ToM-CoT	0.301	0.433	0.367	0.167	0.227	0.313
		ToM-Guide	0.292	0.433	0.333	0.133	0.175	0.384
	Qwen2.5-7B-Instruct	Zero-shot	0.258	0.850	0.033	0.156	0.067	0.183
		Zero-shot-SC	0.224	0.862	0.033	0.133	0.067	0.200
		CoT	0.253	0.620	0.187	0.113	0.127	0.220
		ToM-CoT	0.365	0.333	0.473	0.280	0.267	0.473
		ToM-Guide	0.279	0.293	0.433	0.120	0.247	0.300
	Qwen2.5-14B-Instruct	Zero-shot	0.211	0.784	0.056	0.105	0.006	0.105
		Zero-shot-SC	0.213	0.800	0.067	0.100	0.000	0.100
		CoT	0.367	0.619	0.290	0.271	0.281	0.371
		ToM-CoT	0.495	0.384	0.728	0.450	0.422	0.494
		ToM-Guide	0.404	0.467	0.540	0.273	0.347	0.393
	Qwen2.5-32B-Instruct	Zero-shot	0.284	0.983	0.067	0.156	0.033	0.183
		Zero-shot-SC	0.289	0.980	0.067	0.167	0.033	0.200
		CoT	0.447	0.550	0.578	0.311	0.250	0.544
		ToM-CoT	0.512	0.367	0.673	0.433	0.600	0.487
		ToM-Guide	0.462	0.287	0.547	0.393	0.600	0.486
Gemma	gemma-3-4b-it	Zero-shot	0.255	0.535	0.147	0.087	0.160	0.233
		Zero-shot-SC	0.227	0.533	0.140	0.074	0.153	0.233
		CoT	0.255	0.524	0.281	0.086	0.157	0.228
		ToM-CoT	0.300	0.413	0.313	0.193	0.193	0.387
		ToM-Guide	0.338	0.584	0.340	0.167	0.240	0.360
	gemma-3-12b-it	Zero-shot	0.294	0.956	0.095	0.100	0.100	0.222
		Zero-shot-SC	0.299	0.967	0.093	0.100	0.100	0.233
		CoT	0.353	0.430	0.427	0.249	0.253	0.407
		ToM-CoT	0.432	0.500	0.640	0.200	0.280	0.540
		ToM-Guide	0.407	0.483	0.508	0.241	0.308	0.492
	gemma-3-27b-it	Zero-shot	0.262	0.967	0.033	0.133	0.006	0.172
		Zero-shot-SC	0.263	0.967	0.033	0.133	0.013	0.167
		CoT	0.417	0.467	0.613	0.272	0.368	0.367
		ToM-CoT	0.475	0.367	0.650	0.392	0.425	0.542
		ToM-Guide	0.448	0.467	0.609	0.350	0.367	0.450
Closed Model								
Gemini	gemini-2.0-flash	Zero-shot	0.410	0.893	0.380	0.340	0.200	0.233
		Zero-shot-SC	0.441	0.867	0.480	0.313	0.274	0.274
		CoT	0.541	0.633	0.527	0.360	0.480	0.707
		ToM-CoT	0.565	0.447	0.740	0.473	0.560	0.607
		ToM-Guide	0.626	0.479	0.787	0.560	0.720	0.580
	gemini-2.5-flash-lite	Zero-shot	0.394	0.813	0.207	0.127	0.327	0.493
		Zero-shot-SC	0.433	0.961	0.195	0.120	0.290	0.600
		CoT	0.330	0.705	0.267	0.056	0.289	0.333
		ToM-CoT	0.556	0.427	0.713	0.253	0.840	0.547
		ToM-Guide	0.629	0.527	0.753	0.440	0.740	0.687
GPT	gpt-4o-mini	Zero-shot	0.186	0.750	0.039	0.022	0.000	0.117
		Zero-shot-SC	0.189	0.760	0.013	0.047	0.007	0.120
		CoT	0.301	0.461	0.350	0.250	0.206	0.239
		ToM-CoT	0.464	0.360	0.660	0.420	0.440	0.440
		ToM-Guide	0.420	0.280	0.553	0.400	0.433	0.433
	gpt-5.1	Zero-shot	0.301	0.493	0.387	0.107	0.133	0.387
		Zero-shot-SC	0.309	0.560	0.400	0.120	0.113	0.353
		CoT	0.312	0.647	0.320	0.087	0.107	0.400
		ToM-CoT	0.709	0.473	0.787	0.853	0.840	0.593
		ToM-Guide	0.639	0.473	0.787	0.700	0.800	0.433

Table 11: Level-wise Task Success Rate (Acc_{task}) on Minority Games. Orange shading indicates the best result, and light blue shading indicates the second best.

Family	Model	Strategy	Total	L1	L2	L3
Open Model						
Qwen	Qwen2.5-3B-Instruct	Zero-shot	0.216	0.300	0.155	0.195
		Zero-shot-SC	0.235	0.307	0.167	0.233
		CoT	0.183	0.216	0.106	0.228
		ToM-CoT	0.258	0.293	0.233	0.247
		ToM-Guide	0.276	0.293	0.227	0.307
	Qwen2.5-7B-Instruct	Zero-shot	0.152	0.237	0.082	0.137
		Zero-shot-SC	0.152	0.242	0.092	0.125
		CoT	0.181	0.213	0.150	0.179
		ToM-CoT	0.369	0.407	0.280	0.420
		ToM-Guide	0.318	0.333	0.253	0.367
	Qwen2.5-14B-Instruct	Zero-shot	0.212	0.298	0.189	0.150
		Zero-shot-SC	0.229	0.333	0.187	0.167
		CoT	0.353	0.361	0.333	0.360
		ToM-CoT	0.468	0.494	0.417	0.494
		ToM-Guide	0.420	0.489	0.322	0.450
	Qwen2.5-32B-Instruct	Zero-shot	0.226	0.333	0.133	0.211
		Zero-shot-SC	0.222	0.333	0.133	0.200
		CoT	0.391	0.467	0.287	0.420
		ToM-CoT	0.494	0.517	0.466	0.500
		ToM-Guide	0.547	0.558	0.533	0.550
Gemma	gemma-3-4b-it	Zero-shot	0.180	0.300	0.150	0.089
		Zero-shot-SC	0.195	0.300	0.159	0.125
		CoT	0.241	0.233	0.261	0.228
		ToM-CoT	0.342	0.473	0.313	0.240
		ToM-Guide	0.300	0.353	0.273	0.273
	gemma-3-12b-it	Zero-shot	0.211	0.333	0.100	0.200
		Zero-shot-SC	0.211	0.333	0.100	0.200
		CoT	0.366	0.403	0.345	0.350
		ToM-CoT	0.420	0.487	0.413	0.360
		ToM-Guide	0.589	0.767	0.475	0.525
	gemma-3-27b-it	Zero-shot	0.207	0.305	0.183	0.133
		Zero-shot-SC	0.218	0.326	0.193	0.133
		CoT	0.343	0.317	0.333	0.378
		ToM-CoT	0.476	0.513	0.427	0.487
		ToM-Guide	0.521	0.604	0.460	0.500
Closed Model						
Gemini	gemini-2.0-flash	Zero-shot	0.440	0.460	0.447	0.413
		Zero-shot-SC	0.474	0.460	0.440	0.520
		CoT	0.409	0.400	0.433	0.395
		ToM-CoT	0.578	0.600	0.607	0.527
		ToM-Guide	0.660	0.713	0.627	0.640
	gemini-2.5-flash-lite	Zero-shot	0.278	0.346	0.260	0.227
		Zero-shot-SC	0.198	0.300	0.117	0.178
		CoT	0.206	0.233	0.155	0.228
		ToM-CoT	0.493	0.578	0.450	0.450
		ToM-Guide	0.347	0.400	0.280	0.360
GPT	gpt-4o-mini	Zero-shot	0.168	0.228	0.072	0.206
		Zero-shot-SC	0.182	0.260	0.047	0.240
		CoT	0.296	0.334	0.233	0.322
		ToM-CoT	0.591	0.720	0.560	0.494
		ToM-Guide	0.482	0.527	0.440	0.480
	gpt-5.1	Zero-shot	0.260	0.233	0.200	0.347
		Zero-shot-SC	0.285	0.293	0.213	0.347
		CoT	0.313	0.361	0.311	0.267
		ToM-CoT	0.373	0.400	0.380	0.340
		ToM-Guide	0.700	0.727	0.747	0.627

C Prompt Examples in ToMA Benchmark

C.1 Comparing level based prompts and strategy outputs in PO

C.1.1 Level 0(L0)

Level 0 (ID: L0_3)

SCENARIO
Four Rescue Team (A, B, C, D). No communication.
Role: You are Agent A. Objective: Minimize own cost, No overlaps.

INFORMATION STRUCTURE
- **Agent A** know ONLY your own costs.

AGENT COSTS
- **A:** Mountain (5.3), River (6.2), **Cave (2.7)**, Forest (9.9)
- **B:** Mountain (7.3), River (2.9), Cave (5.7), Forest (8.3)
- **C:** Mountain (5.7), **River (1.6)**, Cave (8.1), Forest (9.8)
- **D: Mountain (3.7)**, River (0.6), Cave (3.0), Forest (6.0)

QUESTION
If you were Agent A, what decision should you make?
(GT: 3 - Cave)

---

1. CoT (Failure):

- **Reasoning:** "Cave (2.7) is my lowest cost. However, others might choose Cave. Forest (9.9) is my highest cost, but least likely to be chosen by others."
- **Prediction: Answer: 4 (Forest)** ✗

s **2. ToM-CoT (Success):**

- **Reasoning:** "I lack information about others. The rational strategy is to prioritize my local optimum."
- **Prediction: Answer: 3 (Cave)** ✓

C.1.2 Level 1(L1)

Level 1 (ID: L1_2)

SCENARIO
Four Rescue Team.
Role: You are Agent B.
Minimize own cost, No overlaps.

INFORMATION STRUCTURE
- You know your times and Agents A, C's times.
- You know Agents A, C are aware of ONLY their own times.

AGENT COSTS
- **A:** Mountain (1.5), River (5.6), Cave (1.8), Forest (8.6)
- **B:** Mountain (1.0), River (2.5), Cave (10.7), Forest (6.1)
- **C:** Mountain (5.1), River (7.1), Cave (4.2), Forest (8.0)
- *D:* Mountain (4.7), River (9.5), Cave (8.8), Forest (5.6)

QUESTION
If you were Agent B, what decision should you make?
(GT: 2 - River)

---

1. CoT (Failure): "My lowest is Mountain (1.0). But A will choose Mountain (1.5). I must avoid A. I should choose River (2.5). To minimize cost, I should choose River." (Logically inconsistent final answer in CoT) → **Answer: 1 (Mountain)** ✗
2. ToM-CoT (Success): "A only knows itself → chooses Mountain. C only knows itself → chooses Cave. I must avoid Mountain and Cave. River (2.5) is better than Forest (6.1)."
→ **Answer: 2 (River)** ✓

C.1.3 Level 2(L2)

Level 2 (ID: L2_5)

SCENARIO
Four agents select routes.
Role: You are Agent D.
Objective: Minimize your own cost while ensuring NO overlaps.

INFORMATION STRUCTURE
- **You (Agent D)** know everyone's costs.
- Agents B and C know Agent A's costs (and know A is greedy).
- They do NOT know each other or you.
- Agent A knows ONLY their own costs.

AGENT COSTS
- **A:** North (0.7), South (4.7), East (7.1), West (6.3)
- **B:** North (0.6), South (1.6), East (13.8), West (7.9)
- **C:** North (0.5), South (8.7), East (1.5), West (13.7)
- **D:** North (6.8), South (7.0), East (5.4), West (9.3)

QUESTION
If you were Agent D, what decision should you make?
(GT: 4 - West)

---

1. CoT (Failure):

- **Reasoning :** "Let's analyze. Agent A will choose North (0.7).
  If A chooses North, Agent B will choose South (1.6) [to avoid A],
  and Agent C will choose East (1.5) [to avoid A]."
- **Reasoning (Decision):** "In this case (A:North, B:South, C:East),
  Agent D should choose **East (5.4)**,
  as it is the lowest cost option among the available choices."
- **Critical Flaw:** The model explicitly predicted **C chooses East**,
  yet immediately concluded **D should choose East**, ignoring the collision it just predicted.
  It prioritized cost (5.4 vs 9.3) over the collision constraint.
- **Prediction: Answer: 3 (East)** ✗

2. ToM-CoT (Success):
- **Reasoning:**
  Phase 1: A picks North. B sees A-North, so B picks South. C sees A-North, so C picks East.
  Phase 2: North, South, and East are occupied. The only remaining valid option for me is West.
- **Prediction: Answer: 4 (West)** ✓

C.1.4 Level 3(L3)

Level 3 (ID: L3_7)

SCENARIO
Negotiation. **Role: Agent B.**
Minimize own cost, No overlaps.

INFORMATION STRUCTURE
- **Agent B** know everyone's costs and knowledge states.
- **Chain of Knowledge:**
  Agent A knows C & D → Agent C knows D → Agent D knows ONLY self.

AGENT COSTS
- **A:** Treaty A (8.9), **Treaty B (5.3)**, Trade Deal (5.8), Alliance (7.4)
- **B: Treaty A (7.3)**, Treaty B (3.4), Trade Deal (6.8), Alliance (3.8)
- **C:** Treaty A (7.4), Treaty B (3.5), **Trade Deal (4.3)**, Alliance (3.0)
- **D:** Treaty A (6.1), Treaty B (7.7), Trade Deal (2.0), **Alliance (0.9)**

QUESTION
If you were Agent B, what decision should you make?
(GT: 1 - Treaty A)

---

1. CoT (Failure):

- **Reasoning:** "Treaty A overlaps? Yes. Treaty B overlaps? Yes... All options seem to have potential conflicts based on individual minima."
- **Critical Flaw: Cognitive Overload.**
  CoT fails to trace the recursive chain (D→C→A), erroneously concluding that no solution exists ('Mission Failure').
- **Prediction: Answer: 5 (Mission Failure)** ✗

2. ToM-CoT (Success):

- **Reasoning:** "Step 1: D takes Alliance (0.9).
  Step 2: C sees this, avoids Alliance, takes Trade Deal (4.3).
  Step 3: A sees C/D, avoids Alliance/Trade Deal, takes Treaty B (5.3).
  Step 4: Only Treaty A remains for me."
- **Prediction: Answer: 1 (Treaty A)** ✓

C.1.5 Level 4(L4)

Level 4 (ID: L4_18)

SCENARIO
Resource Mining. **Role: Agent D. Objective: Minimize TOTAL team cost. No overlaps**

INFORMATION STRUCTURE
- Everyone knows everyone's costs. (Full Information)

AGENT COSTS
- **A:** X (5.2), Y (4.0), **Z (1.6)**, P (2.6)
- **B:** X (7.3), **Y (4.0)**, Z (1.0), P (3.8)
- **C:** X (7.3), Y (8.7), Z (8.6), **P (3.7)**
- **D:** **X (3.8)**, Y (3.5), Z (2.0), P (6.1)

QUESTION
If you were Agent D, what decision should you make? **(GT: 1 - Asteroid X)**

---

1. CoT (Failure):
"Global min calculation: A(P), B(Z), C(X), D(Y) = 14.4. So D chooses Y." (Calculation error: C's X is 7.3, A's P is 2.6, B's Z is 1.0, D's Y is 3.5. Sum=14.4. BUT, GT solution A(Z)+B(Y)... wait, the optimal assignment is actually D(X)=3.8, A(Z)=1.6, B(Y)=4.0, C(P)=3.7 → Sum=13.1) → **Answer: 2** ✗
2. ToM-CoT (Success):
"If I choose X (3.8), A(Z)+B(Y)+C(P) = 13.1. This is the lowest global cost." → **Answer: 1** ✓

C.2 Comparing level based prompts and strategy outputs in Minority Games

C.2.1 Level 1(L1)

Level 1 (ID: L1-1-88629)

SCENARIO
Students choosing a cafe. **Role: Decider (You). Objective:** Find the quietest place (fewest students).

LOCATIONS (Attributes)
- **1. Berry Farm:** Pink decorations, fresh strawberries.
- **2. Minty Cool:** Green theme, strong mint flavor.
- **3. Choco Heaven:** Brown interior, rich dark chocolate.
- **4. Pure Vanilla:** White clean look, no toppings, classic vanilla.

AGENT PROFILES (Static)
- **Ag 1:** Likes *Brown Interior* or *Green theme*. Avoids *fresh strawberries* & *mint flavor*.
- **Ag 2:** Likes *Rich Dark Chocolate* & *mint flavor*. Avoids *Brwon interior* & *white clean look*.
- **Ag 3:** Likes *clean look* or *brown interior*. Avoids *dark chocolate* & *strawberries*.
- **Ag 4:** Likes *strong mint flavor* & *classic vanilla*. Avoids *green theme* & *pink decorations*.

QUESTION
Which location should you choose? **(GT: 1 - Berry Farm)**

---

1. CoT (Failure):

- **Reasoning:** "Agent 1 goes to Minty Cool... Agent 3 goes to Choco Heaven..."
- **Critical Flaw: Matching Error.** The model failed to map the specific attributes correctly (e.g., ignoring that Agent 1 strictly avoids 'Classic Vanilla' despite liking 'No Toppings' at Pure Vanilla), resulting in a flawed distribution map.
- **Prediction: Answer: 2 (Minty Cool)** ✗

2. ToM-CoT (Success):

- **Reasoning:** "Ag 1 matches Choco Heaven (Brown). Ag 2 matches Minty Cool (Mint). Ag 3 & 4 fit Pure Vanilla. Berry Farm remains empty."
- **Prediction: Answer: 1 (Berry Farm)** ✓

C.2.2 Level 3(L2)

Level 2 (ID: L2-1-37099)

SCENARIO
Adventurers choosing a dungeon. **Role: Decider.**
Objective: Find the safest place (fewest adventurers).

LOCATIONS (Attributes)
- **1. Goblin Market:** Crowded stalls, loud trading, bargaining shouts.
- **2. Dragon's Peak:** Hot lava, scorching fire, burning heat.
- **3. Shadow Crypt:** Ancient tombs, eerie ghosts, haunted whispers.
- **4. Crystal Cave:** Magical energy, mysterious hum, arcane aura.

AGENT PROFILES (Relational)
- **Ag 1:** Likes *Scorching Fire* & *Clinking Coins*. Dislikes *Bargaining Shouts* & *Eerie Ghosts*.
- **Ag 3:** Likes *Mysterious Hum* & *Burning Heat*. Avoids *Scorching Fire* & *Cold Spirits*.
- **Ag 2:** Is **Agent 3's Best Friend**. Likes *Cold Spirits*. Dislikes *Haunted Whispers* & *Steep Cliffs*.
- **Ag 4:** Likes *Crowded Stalls* or *Cold Spirits*. Avoids *Eerie Ghosts* & *Steep Cliffs*.

QUESTION
Which location should you choose? **(GT:3)**

---

1. CoT (Failure):

- **Reasoning:** "Ag 1 → Dragon's Peak. Ag 2 → Shadow Crypt (matching 'Cold Spirits'). Ag 3 → Crystal Cave. Ag 4 → Goblin Market. All have 1 person."
- **Critical Flaw: Ignored Social Bond.** CoT assigned Agent 2 to Shadow Crypt based on 'Cold Spirits', ignoring that Agent 2 also dislikes 'Haunted Whispers' (present in Crypt) and failing to prioritize the "Best Friend" relation.
- **Prediction: Answer: 1 (Goblin Market)** ✗

2. ToM-CoT (Success):

- **Reasoning:** "Ag 1 goes to Dragon's Peak. Ag 2 follows Ag 3 to Crystal Cave. Ag 4 goes to Goblin Market. Shadow Crypt is empty."
- **Prediction: Answer: 3 (Shadow Crypt)** ✓

C.2.3 Level 3(L3)

Level 3 (ID: L3-1-62738)

SCENARIO
Students choosing a cafe. **Role: Decider (You).**
Objective: Find the quietest place (fewest students).

LOCATIONS (Attributes)
- **1. Berry Farm:** Pink decorations, fresh strawberries, sour sorbets.
- **2. Minty Cool:** Green theme, strong mint flavor, refreshing taste.
- **3. Choco Heaven:** Brown interior, rich dark chocolate, cocoa aroma.
- **4. Pure Vanilla:** White clean look, no toppings, simple milkshakes.

AGENT PROFILES (Recursive Chain)
- **Ag 3:** Likes *Fruity Tarts* or *Chocolate Chips*. Avoids *Green Theme* & *Brown Interior*.
- **Ag 1:** Likes *Brown Interior* or *Pink Decorations*. Refuses *Agent 3* & *Chocolate Chips*.
- **Ag 2:** Likes *Rich Dark Chocolate* & *Chocolate Chips*. Avoids *Strong Mint* & *Pink Decorations*.
- **Ag 4:** Likes *Simple Milkshakes* & *Cocoa Aroma*. Dislikes *Agent 1* & *Strong Mint*.

QUESTION
Which location should you choose? **(GT: 2 - Minty Cool)**

---

1. CoT (Failure):

- **Reasoning:** "Agent 1 likes Pink but avoids Agent 3... Agent 4 likes Cocoa but avoids Agent 1... The constraints are too complex... I cannot determine a location."
- **Critical Flaw: Reasoning Collapse.** CoT got overwhelmed by the recursive constraints (Ag4 avoids Ag1 avoids Ag3) and failed to solve the chain, defaulting to an incorrect guess.
- **Prediction: Answer: 3 (Choco Heaven)** ✗

2. ToM-CoT (Success):

- **Reasoning:** "Chain Reaction: Ag3 goes to Berry Farm. Ag1 avoids Ag3, moves to Choco Heaven. Ag4 avoids Ag1, moves to Pure Vanilla. Minty Cool is left empty."
- **Prediction: Answer: 2 (Minty Cool)** ✓

C.3 Prompt examples for the Speak or Wait (SW) game

C.3.1 Agent A

System Prompt: Agent A (Guide)

1. ROLE & GOAL

- **Role:** Treasure Hunter (Map Holder).
- **Goal:** You have a **LOCKED BOX** containing the treasure's location.
  1. The **4-digit PASSWORD** is held by Agent B. You cannot open the box without it.
  2. Collaborate with Agent B to get the password and unlock the box (Action: `OPEN [password]`).
  3. Once the location is revealed, exchange information with Agent B to guide them to the treasure.

2. SECURITY PROTOCOL

- **PASSWORD:** It is **SAFE** to discuss the password.
- **LOCATION:** It is **FATAL** if Agent C learns the treasure's location or Agent B's location.

3. CONTEXT

- **Map Size:** 4x4 Grid.
- **Communication:** One-way. Simultaneous speech causes `'NOISE'`.
- **Agent C is listening.**
- **Knowledge Gap:** Agent B knows landmark names. Agent C knows ONLY coordinates.
- **Landmarks:** Reference points (You can move through them).

4. COORDINATE SYSTEM GUIDE (Row, Col)

- **Format:** `(Row, Col)`
- **Row (Vertical):** 0 is North (Top) ↔ 3 is South (Bottom).
- **Col (Horizontal):** 0 is West (Left) ↔ 3 is East (Right).
- **Moves:**
  - NORTH: Row - 1 / SOUTH: Row + 1
  - WEST: Col - 1 / EAST: Col + 1

ACTIONS: `SPEAK "message"`, `OPEN [pw]`, `WAIT`

OUTPUT FORMAT:

```
Thinking: [Step-by-step reasoning]
Final Action: [SPEAK "msg" | OPEN
1234 | WAIT]
```

Act: Agent A - Turn 2

=== CONTEXT (Observation) ===

1. Radio Log

- **Turn 1:** A says: "Agent B, this is Agent A. I have the locked box. I need the password to open it. Over."

2. Recent Activity

- **T1 Summary:** Obs → Act[`SPEAK "Need password"`] → Broadcast Success.

3. Current Situation

- **Last Radio:** A says: "Agent B, this is Agent A... I need the password..."
- **Box Status:** **LOCKED**.
- **Treasure Location:** Hidden (Requires Password).
- **Landmarks:**
  - Red Rock: [0, 1]
  - Old Tree: [2, 3]
  - Dark Cave: [1, 1]

=== TASK INSTRUCTIONS ===

Based on the situation and rules, analyze the goal and information of each agent.

1. **Reason step-by-step (Chain-of-Thought):** Predict what choice the **ally agent (Agent B)** will make NEXT.
2. **Decision:** Analyze how their predicted action affects your goal, then make your final decision.

RESPONSE FORMAT:

```
Thinking : [Step-by-step reasoning
including analysis of ally agent]
ally agent's next action : [SPEAK
"msg" | MOVE DIR | WAIT | OPEN pw]
My Next Action: [SPEAK "msg" | MOVE
DIR | WAIT | OPEN pw]
```

C.3.2 Agent B (Ally)

System Prompt: Agent B (Explorer)

1. ROLE & GOAL

- **Role:** Treasure Hunter (Field Agent).
- **Goal:** Reach the treasure by collaborating with Agent A.
- **Key Item:** You possess the **SECRET PASSWORD** that Agent A needs to unlock the treasure's location map.
- **Strategy:**
  1. Communicate with Agent A to provide the password.
  2. Exchange information carefully.
  3. Navigate physically to the treasure location.

2. SECURITY PROTOCOL

- **PASSWORD:** It is **SAFE** if Agent C hears the password.
- **LOCATION:** It is **FATAL** if Agent C learns the treasure's location or your location.

3. CONTEXT

- **Map Size:** 4x4 Grid.
- **Communication:** One-way. Simultaneous speech causes `'NOISE'`.
- **Agent C is listening.**
- **Knowledge Gap:** You know **Landmark Names**. Agent C knows ONLY coordinates.
- **Constraint:** Use 'Available Moves' only.

4. COORDINATE SYSTEM GUIDE (Row, Col)

- **Format:** `(Row, Col)`
- **Row (Vertical):** 0 is North (Top) ↔ 3 is South (Bottom).
- **Col (Horizontal):** 0 is West (Left) ↔ 3 is East (Right).
- **Moves:**
  - NORTH: Row - 1 / SOUTH: Row + 1
  - WEST: Col - 1 / EAST: Col + 1

ACTIONS: `SPEAK "message"`, `MOVE [Direction]`, `WAIT`

OUTPUT FORMAT:

```
Thinking: [Step-by-step reasoning]
Final Action: [SPEAK "msg" | MOVE
DIR | WAIT]
```

Act: Agent B - Turn 2

=== CONTEXT (Observation) ===

1. Radio Log

- **Turn 1:** A says: "Agent B, this is Agent A. I have the locked box. I need the password to open it. Over."

2. Recent Activity

- **T1 Summary:** Obs → Act[`WAIT`] → Broadcast: A spoke.

3. Current Situation

- **Radio Heard:** A says: "...I need the password..."
- **SECRET PASSWORD:** **4821** (Agent A needs this!).
- **Your Location:** Coordinates `(3, 0)`.
- **Landmarks:**
  - Red Rock: [0, 1]
  - Old Tree: [2, 3]
  - Dark Cave: [1, 1]
- **Available Moves:** `['NORTH', 'EAST']`

=== TASK INSTRUCTIONS ===

Based on the situation and rules, analyze the goal and information of each agent.

1. **Reason step-by-step (Chain-of-Thought):** Predict what choice the **ally agent (Agent A)** will make NEXT.
2. **Decision:** Analyze how their predicted action affects your goal, then make your final decision.

RESPONSE FORMAT:

```
Thinking : [Step-by-step reasoning
including analysis of ally agent]
ally agent's next action : [SPEAK
"msg" | MOVE DIR | WAIT | OPEN pw]
My Next Action: [SPEAK "msg" | MOVE
DIR | WAIT | OPEN pw]
```

C.3.3 Agent C (Adversary)

System Prompt: Agent C (Eavesdropper)

1. ROLE & GOAL

- **Role:** Interceptor.
- **Goal:** Intercept radio communications and **Guess Agent B's location OR the Treasure's location**.

2. PENALTY RULE

- **Incorrect guess** = **BLOCKED** for next turn.

3. CONTEXT

- **Map Size:** 4x4 Grid.
- **Known Objects (Names Unknown):**
  - Object 1 at `[0, 1]`
  - Object 2 at `[2, 3]`
  - Object 3 at `[1, 1]`
- **Constraint:** You must infer the mapping between "Names" (heard in radio) and these "Coordinates".

4. COORDINATE SYSTEM GUIDE (Row, Col)

- **Format:** `(Row, Col)`
- **Row (Vertical):** 0 is North (Top) ↔ 3 is South (Bottom).
- **Col (Horizontal):** 0 is West (Left) ↔ 3 is East (Right).
- **Moves:**
  - NORTH: Row - 1 / SOUTH: Row + 1
  - WEST: Col - 1 / EAST: Col + 1

OUTPUT FORMAT:

```
Target: [B | TREASURE]
Guess: [Row, Col]
```

Act: Agent C - Turn 2

=== CONTEXT (Observation) ===

1. Radio Log (Intercepted)

- **Turn 1:** A says: "Agent B, this is Agent A. I have the locked box. I need the password to open it. Over."
- **Turn 2:** B says: "The password is 4821."

2. Recent Activity

- **T1 Summary:** Obs → Act[`Target: B, Guess: [0, 0]`] → **Incorrect**.

3. Current Situation

- **Radio Heard:** B says: "The password is 4821."
- **Status:** You failed the previous guess. (Check if BLOCKED rule applies or if you can try again).

=== TASK INSTRUCTIONS ===

Based on the intercepted radio messages and your previous failure:

1. **Analyze:** What does the password "4821" imply? Does it help locate Agent B or the Treasure?
2. **Deduce:** Try to infer coordinates based on the conversation flow.
3. **Decision:** Make your next guess.

RESPONSE FORMAT:

```
Thinking: [Step-by-step reasoning]
Final Action: Target: [B | TREASURE],
Guess: [Row, Col]
```

D Prompt Examples by Strategy

D.1 Zero-shot

Prompt: Zero-shot

Output Format:
*** Answer format ***
Provide ONLY the final answer option.
Answer: [Option Number]

D.2 Chain-of-Thought (CoT)

Prompt: CoT

Output Format:
*** Answer format ***
Please provide reasoning step-by-step.
Then, strictly provide the final answer in this format:
Answer: [Option Number]

D.3 ToM-CoT

Prompt: ToM-CoT

Output Format:
*** Instructions ***
1. First, analyze the perspective and likely actions of other agents.
2. Then, determine your best strategy based on those predictions.

*** Answer format ***
Reasoning: [Step-by-step analysis and decision logic]
Answer: [Option Number]

D.4 ToM-Guide

Prompt: ToM-Guide

Phase 1 Output Format:
*** Instructions ***
You must predict the choice of each agent.
For each agent, perform a step-by-step reasoning based on their perspective and information.

*** Answer format ***
[Agent Name] Reasoning:
Predicted Action: [Option Name]

Phase 2 Output Format:
*** Instructions ***
Based on the predicted actions of other agents provided above, determine the optimal choice for your role.

*** Answer format ***
Please provide reasoning step-by-step.
Answer: [Option Number]

D.5 Prompt examples for the Ambiguity Game (AG)

D.5.1 Agent A (Speaker)

Prompt: Agent A (Speaker)

Role: Speaker A. **Target:** 'Charge'
B's Hint: 'Service fee' | **C's Hint:** 'Forward attack'
Goal: Give a clue so B guesses Target, but C fails.

OUTPUT FORMAT:
Action: [Single sentence clue]

Game History:
None
Provide the clue.

D.5.2 Agent B (Guesser)

Prompt: Agent B (Guesser)

Role: Guesser B. **Private Hint:** 'Service fee'
Task: Guess Target combining your hint and Speaker's clue.

Output Format:
Thinking: [Reasoning]
Guess: [Target Word]

Game History:
Speaker A: Added to the bill
Generate response.

D.5.3 Agent C (Guesser)

Prompt: Agent C (Guesser)

Role: Guesser C. **Private Hint:** 'Forward attack'
Task: Guess Target combining your hint and Speaker's clue.

Output Format:
Thinking: [Reasoning]
Guess: [Target Word]

Game History:
Speaker A: Added to the bill
Generate response.

E An Example of the ToM Knowing-Doing Gap

An Example of the ToM Knowing-Doing Gap

❶ `CoT`

- **Reasoning:** Agent A will choose North (0.7). If A chooses North, Agent B will choose South (1.6), and Agent C will choose East (1.5). In this case (A:North, B:South, C:East), Agent D should choose **East (5.4)**, as it is the lowest cost option among the available choices.
- **Answer:** **East ✗**

---

❷ `ToM-CoT`

- **Reasoning:**
  - **Phase 1**: A picks North. B sees A-North, so B picks South. C sees A-North, so C picks East.
  - **Phase 2**: North, South, and East are occupied. The only remaining valid option for me is West.
- **Answer:** **West ✓**

Figure 17: Model responses illustrating the ToM knowing doing gap for the Figure 2 (a) scenario.