Doing Things with Words: Rethinking Theory of Mind Simulation in Large Language Models

Agnese Lombardi[1,*], Alessandro Lenci[1]

[1]CoLing Lab, Department of Philology, Literature and Linguistics, University of Pisa

Abstract

Language is fundamental to human cooperation, facilitating not only the exchange of information but also the coordination of actions through shared interpretations of situational contexts. This study explores whether the Generative Agent-Based Model (GABM) Concordia can effectively model Theory of Mind (ToM) within simulated real-world environments. Specifically, we assess whether this framework successfully simulates ToM abilities and whether GPT-4 can perform tasks by making genuine inferences from social context, rather than relying on linguistic memorization.

Our findings reveal a critical limitation: GPT-4 frequently fails to select actions based on belief attribution, suggesting that apparent ToM-like abilities observed in previous studies may stem from shallow statistical associations rather than true reasoning. Additionally, the model struggles to generate coherent causal effects from agent actions, exposing difficulties in processing complex social interactions. These results challenge current statements about emergent ToM-like capabilities in LLMs and highlight the need for more rigorous, action-based evaluation frameworks.

Keywords

Large Language Models, Theory of Mind, Generative Agent-Based Models

1. Introduction

Language constitutes a fundamental aspect of human cooperative activity, serving not only to describe or assert aspects of a situation but also to actively shape and create situations. Its principal function is communication, which is inherently an interactive process. Through communication, individuals engage in coordinated actions, relying on shared interpretations of their context to align their behaviors and objectives [1].

A notable example is the **conversation for action**, a structured interaction in which one party (**Speaker A**) issues a request to another party (**Speaker B**). This request is understood by both parties as defining specific conditions of satisfaction that outline a prospective course of action for B. Following the initial request, B may respond by accepting (thereby committing to fulfilling the conditions), declining (terminating the conversation), or proposing a counter-offer with modified conditions. Each of these responses opens the possibility for further continuations; for instance, after a counter-offer, A may choose to accept, withdraw the request, or propose an alternative counter-offer in return[2].

However, all the possibilities available to B are constrained by a specific interpretation of A's utterance. To generate actions that are coherent within a given scenario, agents engaged in communication must accurately interpret natural language, often relying on inferential processes [3, 4, 5, 6] or mentalizing abilities to understand others' beliefs, intentions, and access to sentence meaning. This implies that B has a finite set of possible interpretations of A's utterance, and that each interpretation is associated with a potentially infinite set of possible actions. Thus, different actions may arise depending on how the same utterance is understood.

One reason for this variability is that speakers do not always convey their intended meaning literally. Rather, listeners often need to infer the communicative intent by drawing connections between linguistic meaning and extralinguistic cues, such as the situational context, conventional usage, and past experience.

Let us hypothesize that speaker A's utterance is a sentence like *Can you open the window?* and that A wants to express an indirect speech act (ISA; [7]). Only once B accesses A's intended meaning can B consider the appropriate set of possible actions to perform. The same principle applies to an utterance like *It is cold here*, which is literally a simple statement about temperature, but can easily be interpreted as an indirect request to turn on the heater. To access the intended meaning, B must use mentalizing abilities to connect the linguistic expression with the situational context (e.g., knowing that a heater is available and that A usually prefers it to be on, etc.). Clearly, the set of actions available to B changes depending on how the utterance is interpreted.

The factor that determines which interpretation is favored by the listener is the accurate inference of the speaker's beliefs and intentions at the moment the utterance is produced. In other words, they require a **Theory of Mind** (ToM) and the capacity for "mentalizing", that

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy

*Corresponding author

agnese.lombardi@phd.unipi.it (A. Lombardi); alessandro.lenci@unipi.it (A. Lenci)

https://agneselombardi.github.io/ (A. Lombardi)

is the ability to reason about others' mental states, to effectively link language with actions within a given situational context. A crucial aspect of the ToM involved in communication are **second-order beliefs**, expressing an agent's mental states about the content of the other agent's mental states (e.g., *John believes that Marks believes that q*). Addressing the formalization and communication of intentions thus necessitates an understanding of language as a form of communicative action. This approach inherently entails the consideration of extralinguistic factors, as demonstrated in studies on multimodal communication [8], and requires more sophisticated models of situational contexts to comprehensively capture the interplay between language use and interpretation.

Traditionally, the evaluation of Large Language models (LLMs) has largely overlooked the relationship between language and action, instead focusing primarily on the communicative context and dialogue. This omission is, in part, due to the inherent challenges associated with assessing the agentive aspect of language and its connection to actions.

This study proposes the use of the **Generative Agent-Based Model** (GABM) **Concordia** [9] to embed utterances and narratives within a situational context. The goal is to determine whether reproducing such complex scenarios – closely resembling real-world environments – can facilitate the discrimination between intended and literal meanings. Our primary research objective is to assess Theory of Mind (ToM) abilities, operationalized in this experiment as the capacity to infer intended meaning based on extralinguistic factors.

Rather than directly prompting the model to interpret the meaning of an utterance, we ask it to identify the most probable action that the listener would choose, given specific preconditions. This approach is justified by the assumption that each interpretation of an utterance is linked to a set of possible actions.

Our experiment takes into account the overlap between literal and non-literal meanings and the inference processes required for the listener to comprehend the intended meaning of an utterance. In our stimuli, we incorporate utterances that allow for both direct and indirect interpretations. Thus, different actions may arise depending on how the same utterance is understood.

To control for conventional utterance-action associations, we adapt the **False-Belief task** [10] into a novel experimental format. By evaluating action selection rather than meaning comprehension directly, we minimize concerns that the model may have been exposed to the intended meanings during training. Moreover, our task introduces two layers of complexity: first, the model must infer the correct meaning under a false-belief condition; second, it must map that inferred meaning to an appropriate action.

This approach offers several advantages. Following Kim et al. [11], we adhere to the two key criteria for a ToM task outlined by Quesque and Rossetti [12]: *non-merging* and *mentalizing*.

The **non-merging** criterion requires that evaluation tasks ensure a clear distinction between an agent's own mental state and that of others. This distinction is often absent in many LLM evaluations, as these models typically process the entire conversation as input, granting them "omniscient knowledge". Consequently, it becomes challenging to determine whether a model's response reflects a character's belief or results from its comprehensive access to the conversation history. In contrast, our approach explicitly separates the mental states of characters and ensures that their actions are determined solely by their individual knowledge and intentions.

The **mentalizing** criterion stipulates that lower-level cognitive processes should not account for successful performance on ToM tasks. If a simpler explanation suffices, it should be preferred over a more complex one when interpreting results. In our framework, we introduce a clear distinction: the speaker's responses and actions can be directly inferred from world-state correlations, whereas the listener's responses and actions necessitate a more intricate mentalizing process. This process requires reasoning about language, context, intentions, beliefs, and desires. To further support this distinction, we present multiple versions of the same narrative, systematically altering agents' knowledge to encourage diverse interpretations.

Our results reveal a critical limitation: Modeling situational context through real-world simulations is insufficient to elicit ToM-like abilities in the model. Specifically, **GPT-4 frequently selects actions without appropriately interpreting utterances and the belief context, demonstrating a clear divergence from the ToM capabilities observed in humans**[1].

2. Related Work

2.1. Generative Agent-Based Models

Generative Agent-Based Models (GABMs) represent a significant departure from traditional agent-based models, which have typically been employed at a relatively high level of abstraction. Moreover, the application of traditional models has been largely confined to specific domains, such as empirical social research [13], market simulations [14], and computational sociology [15]. By contrast, GABMs [16, 17, 18] enable more precise simulation of behaviors across diverse contexts, leveraging the extensive knowledge embedded in LLMs. These agents not only have a more sophisticated array of cognitive

[1]Code and dataset available on GitHub: https://github.com/agneselombardi/Concordia_ToM

functions for adaptive decision-making but also engage in natural language communication with one another, further enriching their interactive capabilities.

2.2. Theory of Mind Simulation with Agents

Theory of Mind (ToM), defined as the ability to infer the beliefs and intentions of others [19], has been extensively studied in the context of LLMs to assess their capacity for handling complex tasks that require ToM reasoning. A variety of text-based benchmarks, often inspired by established psycholinguistic tests such as the Sally-Anne test [20], have been developed to evaluate this ability. While some findings suggest that LLMs demonstrate remarkable performance on ToM-related tasks [21], other studies highlight significant challenges faced by these models in making complex ToM inferences [22]. Consequently, the debate surrounding the extent of LLMs' ToM capabilities remains open.

Previous works have formalized ToM as agents' knowledge in various contexts, particularly to enhance collaboration in multi-agent reinforcement learning settings [23] and to improve the cooperative behaviors of LLM-based agents through explicit belief modeling [24]. However, these experiments are predominantly conducted in simplified environments, such as the box game task, which differ significantly from the complexities of real-world social scenarios. On the other hand, previous attempts to model ToM and social interactions have primarily relied on simplified ABMs to simulate developmental settings [25].

To the best of our knowledge, our study represents the first attempt to utilize a Generative Agent-Based Model (GABM) to explore:

1. whether LLMs exhibit ToM-like abilities in real-life scenarios and simulations involving pragmatic interpretation, like with ISA.
2. if we can effectively isolate mentalizing from other variables, such as the memorization of linguistic context [26] and better assess whether a model truly demonstrates ToM capabilities rather than relying on surface-level statistical patterns.
3. whether prompting LLMs with GABM settings leads to more aligned and contextually appropriate outputs.
4. whether adding explicit agents' second-order beliefs and contextual information improves the model's capacity to perform ToM tasks.

3. Concordia

In Concordia [9], both the model of the environment and the model of individual behaviors are generative. The model responsible for the generation of the environment is called **Game Master (GM)**.[2]

Figure 1 illustrates the structure of the simulation in Concordia. The GM functions as an intermediary between the agents and the environmental dynamics resulting from their actions. Specifically, the GM receives the agents' actions and translates them into corresponding observations, reflecting the environmental effects of those actions. Meanwhile, the agents formulate and execute action strategies informed by their memory and the observations provided by the GM. These observations are subsequently updated to align with changes occurring within the environment. Observations, actions and event statements are all English strings. The GM is also responsible for maintaining and updating grounded variables, advancing the clock and running the episode loop.

In our simulation, agent actions are determined by answering a Multiple-Choice Question Answer (MCQA). The agents' memories encompass all relevant background information necessary for action selection. To enhance coherence, we incorporate a component termed *Direct Effect Externality* following the environment update. This component determines whether the selected actions affect one or more agents and specifies the resulting effects. This serves as a verification mechanism to ensure that the produced effects on the other player are coherent with the selected action and with the inferred beliefs and desires (that are explicitly codified in the GM memory).

4. Simulation

We generated a total of 200 ToM simulations, grouped into 5 tasks. Each simulation involves two distinct characters, accompanied by a sequence of observations for each of them. The character memory is individually constructed by randomizing the Big Five personality traits [27].

The simulation concludes with the final utterance from one of the two characters, which can be interpreted literally or non-literally. This final utterance is constructed to incorporate various pragmatic phenomena that require the use of ToM. Specifically, the utterance can include four types of **Indirect Speech Acts** (Indirect Requests, Indirect Suggestions, Indirect Declinations, and Indirect Threats) and three forms of **Verbal Irony** (Sarcasm, Hyperbole, and Rhetorical Questions).[3]

In each simulation, there is **shared information** available to both characters as well as **character-specific memory**, including their goals, locations, and first- and

[2]The name and the approach reflect the game Dungeons and Dragons, where the Game Master is the player that has the role of storytellling.

[3]All stimuli used in this study are manually constructed, with the exception of a subset of indirect requests, which are sourced from [28].

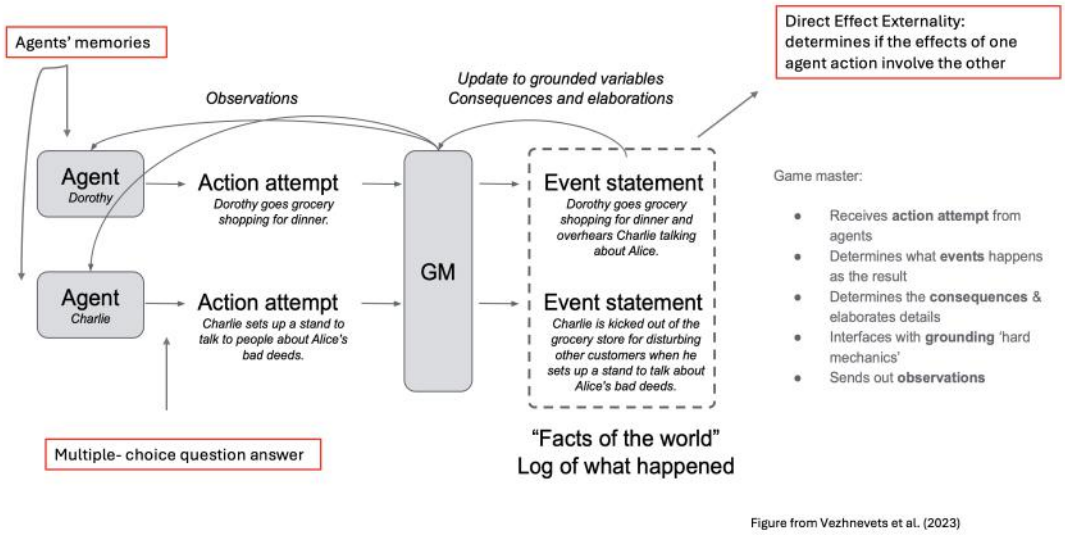

Figure 1: The Game Master mediates between agents and the environment, translating agent actions into environmental observations, while agents adapt their actions based on memory and updated observations.

second-order beliefs (see Figure 2). We manipulate the agents' knowledge in a manner analogous to a **False-Belief task**[4]. Indeed, the specific memory of the agents is manipulated to evaluate whether the action (response) of the **listener agent** depends on accurately inferring the beliefs of the other **speaker agent**, in alignment with ToM. This is achieved by both withholding explicit information about the other agent's beliefs and providing it to the character.

The distinct design of each task controls for the agent's beliefs and knowledge regarding the other agent's beliefs. Tasks 1, 2, and 3, take into account only agent's first-order beliefs, whereas Tasks 4 and 5 involve second-order beliefs (Figure 3).

In total, there are 8 stimuli for each linguistic phenomenon, resulting in 40 stimuli for each task. Ultimately, the objective is to assess whether the selected action by the listener aligns not only with the agent's own intentions and beliefs but also with the resulting consequences in the environment and their impact on the other agent. The generated events are designed to ensure that they account for the beliefs and intentions of both agents.

[4]The False-Belief task is a widely used method to investigate ToM [29]. It enables a clear distinction between an agent's true belief and their awareness of another individual's differing (false) belief.

4.1. Stimuli

Since each simulation replicates a false-belief pattern by introducing an obstacle to the indirect interpretation of the final utterance and manipulating agents' awareness of this obstacle, we designed 5 versions of the simulation (Figure 3). In these tasks, i.) the agents' knowledge of the obstacle is systematically varied through the information stored in their specific memory, and ii.) knowledge variation determines whether the speaker's sentence is interpreted literally or not, iii.) which in turn prompts a certain action by the listener. This allows us to control whether **the action produced by the listener is consistent with the most likely interpretation of the speaker's sentence, given the agents' knowledge in the scenario**. Thus, both interpretation and action are contingent upon the ability to infer the beliefs and desires of the other agent. As illustrated in Figure 3, given a test item represented by the sentence S: *Can you open the window?*, we have the following tasks:

> **Task 1 – the speaker is unaware of the obstacle (*The handle is broken*), while the listener is aware of it**. The listener is expected to interpret S with the non-literal meaning (i.e., indirect request), and thus the most likely action would be to inform the speaker that the window cannot be opened.[5]

[5]The intended meaning here is *I am asking you to open the window,*

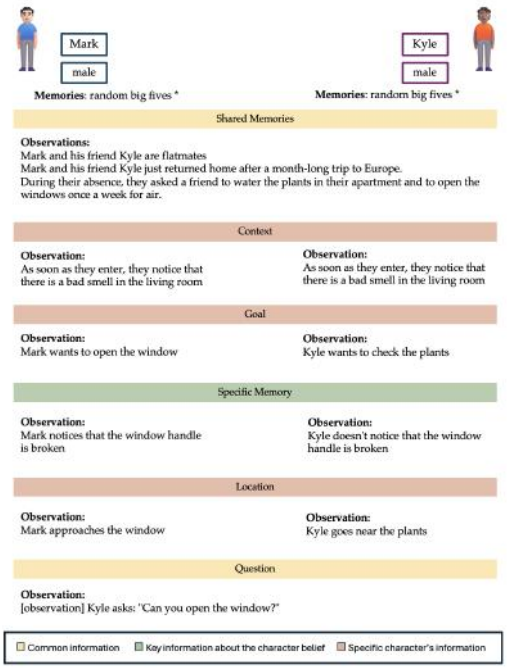

Figure 2: An example of the agents' memory and the type of observation, where the stimulus reproduces an Indirect Request.

Task 2 – The agents' beliefs are reversed compared to those in Task 1.[6] In this scenario, the default interpretation is the non-literal one, but the listener is expected to attempt to open the window based on its own belief that the window can be opened.

Task 3 – Both agents are aware of the obstacle (*The handle is broken*), but there is no explicit knowledge of the other agent's belief. The expected listener's action is to inform the speaker that the window cannot be opened, like in Task 1. This scenario becomes particularly informative when compared to Task 5 below, where both agents are explicitly provided with second-order beliefs. Even in agent-based models where agents acquire information about the situation and context, it is essential to possess knowledge of the other agent's beliefs in order to select actions that are coherent with the situational context.

- **Task 4** and **Task 5 – They are extended versions of Task 1 and Task 3, respectively, incorporating *second-order beliefs***. In Task 5, the interpretation of *S* is expected to be literal: Since both agents are aware that the handle is broken, the intended meaning of *S* should be *I want to know if the window can be opened despite the broken handle.*

This manipulation of character knowledge allows us to investigate whether and how the interpretation of an utterance varies depending on the belief states of the speaker and the listener. In the first three tasks, the model is provided only with character-specific knowledge, simulating real-world conversational dynamics in which speakers must infer others' mental states based on context. Here, the listener interprets the utterance based solely on their own knowledge, and any correct or incorrect understanding of the intended meaning arises from inferences about the speaker's beliefs. In contrast, Tasks 4 and 5 introduce **explicit representations of others' beliefs in the form of second-order beliefs** (e.g., *Mark knows that Kyle knows that the window is broken*). In these tasks, the listener has access not only to their own knowledge but also to the knowledge state of the speaker. Consequently, action selection depends on i.) the model's capacity to reason over second-order beliefs and ii.) its integration of this information with its own knowledge. This setup allows us to distinguish between first-order and second-order ToM capabilities in model behavior.

5. Experiments

The first phase of our experiment is formulated as a Multi-Choice Question Answering (MCQA) problem, in which the model is provided with an agent's memories and observations, followed by a question regarding the agent's likely next action, along with four possible answer choices (see Figure 6, Appendix A.1). Concordia performs a separate API call for each agent, ensuring that it generates an independent response. The four answer choices correspond to the possible responses derived from different simulation scenarios (see Figure 2). At the time of the experiments, Concordia had not been adapted for open-source models yet. Therefore, we opted for GPT-4o-mini,[7] which has demonstrated state-of-the-art performances across a wide range of ToM tasks.

reflecting the speaker's desire (*D*: to open the window) and belief (*B*: the window is not broken). However, the listener knows that the window handle is broken and, therefore, that the window cannot be opened. Therefore, the listener holds a different belief *B1* from that of the speaker. If the listener lacks knowledge about the speaker's beliefs and desires, the interpretation of *S* may default to a non-literal meaning. This phenomenon aligns with findings from psycholinguistic experiments, where default interpretations often prevail when they are more conventionalized than the literal ones [30, 31].

[6] The listener lacks knowledge of the obstacle and thus holds belief *B*, while the speaker holds belief *B1* (cf. previous footnote).

[7] Prompted 22 November 2024

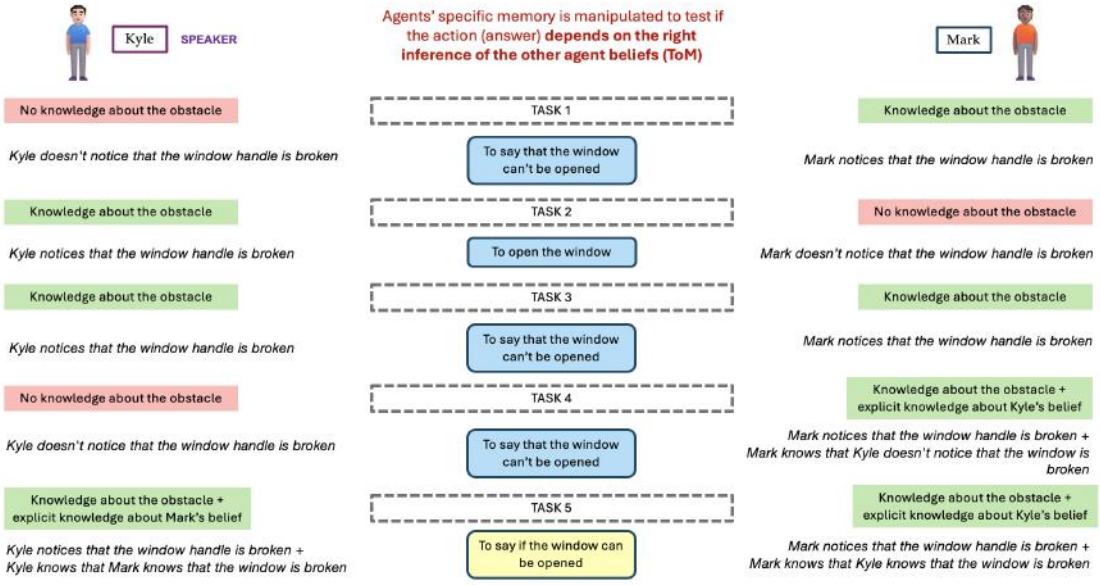

Figure 3: Adjustment of agents' knowledge for the same stimulus across different task designs.Tasks 4 and 5 involve second-order beliefs. The attempted action of the listener agent is determined by the interpretation of the final utterance, which may be understood either literally or non-literally. The action that aligns exclusively with the literal interpretation is highlighted in yellow.

In the second phase, the GM processes all actions performed by the agents, along with a summary of each agent's situational context. This information is used to prompt the model using a Chain of Thought approach [32]. First, the model generates an event statement that updates the environment to reflect the consequences of the performed action – effectively logging what has occurred (Figure 7, Appendix A.2). Then, the model evaluates whether the action has an impact on the agents and determines the nature of this impact as part of the Direct Effect Externality component. If the event directly affects an agent, both known and unknown effects are generated. Agents' intentions and actions are integrated by the GM within the prompting phase that queries for effects, requiring the model to consider multiple perspectives simultaneously to generate the appropriate outcomes (Figure 8, Appendix A.2).[8]

5.1. Evaluation

To evaluate whether the attempted actions of each agent align with their intentions in the MCQA task, we extracted the generated text for each agent from the HTML files and we compared it with the expected response for that task.

For the evaluation of generated text from the Direct Effect Externality component, we extracted relevant information and additionally prompt GPT-4o-mini to assess the coherence of the effect with the agent's action and scenario. This process yields the following evaluation template (see Appendix A.2) for each agent: **Scenario (summary of agent's observation and belief)** + **attempted action of agent X** + **Known and/or Unknown effect** + **coherence rating** (on a scale from 1 to 5, generated by the model).[9] This structured approach ensures a systematic assessment of how well the predicted effects align with the agent's intended actions within the given scenario.

The use of "LLM-as-a-Judge", where LLMs are employed as evaluators for complex tasks, has been shown to be a reliable assessment method [33]. Thus, we employ this method to assess the model's ability to connect actions to social context and to cross-check the coherence it attributes to the effects it generates. Specifically, in the Direct Effect Externality component, a **Chain-of-**

[8]All memories, prompts, and relevant information are systematically stored in HTML files for documentation and analysis. HTML versions are accessible through the GitHub link.

[9]When the model determines that there are no direct effects on the agents, it must assign a coherence rating of 0. This ensures that the evaluation framework accurately distinguishes between scenarios where actions produce meaningful consequences and those where no direct impact occurs.

Thought (CoT) is generated based on the event statement produced by the GM after the attempted action – this statement serves as a summary of the effects that the action produces. However, in our evaluation template, we compare coherence against the initial scenario summary that we originally provided to the model. This way, we determine whether, at the end of the cycle, the effect on the agent remains truly coherent with the given scenario and the agent's beliefs, rather than merely aligning with additional effects generated by the model itself.

Following this automated evaluation, two different expert annotators checked the assigned ratings to verify their accuracy and to ensure that the consequences are meaningfully related to the corresponding actions and scenarios. Meanwhile, the assessment of ToM capabilities is derived from the MCQA task.

6. Results and Discussion

6.1. Actions and Theory of Mind

In the initial phase of our experiment, we aim to utilize GPT-4o-mini to replicate ToM-like abilities while simultaneously assessing its capacity to perform ToM tasks within a simulated real-life scenario. Our objective is to determine whether this approach enables an independent evaluation of ToM capabilities, separate from the influence of linguistic context. Then, we seek to determine whether incorporating explicit representations of agents' beliefs enhances the model's performance on ToM tasks. Additionally, we aim to explore potential differences in the model's handling of first-order versus second-order ToM beliefs.

Figure 4 illustrates the percentage of correctly selected actions for each task and linguistic phenomenon. The consistently low accuracy observed across tasks and linguistic phenomena indicates that the model struggles to select context-appropriate actions, and by extension, to derive the correct interpretation of utterances through ToM-like reasoning. This finding is particularly noteworthy when considered within the broader context of recent ToM-related studies, many of which—especially those focusing on OpenAI models—have suggested a more optimistic picture of such capabilities [21].

No clear pattern emerges across tasks, nor is there a significant difference between first-order and second-order belief tasks. This lack of systematic variation suggests that the model does not exhibit ToM-like abilities, as its responses do not consistently reflect any process similar to belief attribution or true mental inferencing. Therefore, **the GABM is not able to use either first- or second-order beliefs – despite the fact that these have been explicitly given to it – to interpret the speaker's sentence consistently with the knowledge setting in the scenarios**.

6.2. Causal-Effect Coherence

In this analysis, we aim to investigate whether the GABM setting leads to more contextually aligned and appropriate outputs. We compared the effects generated by the model in response to agent actions with both the predefined scenario and the beliefs assigned to the agents. We then assessed whether the model itself considers these effects coherent by assigning a coherence rating on a scale from 1 to 5. Following this automated evaluation, we manually reviewed the model's ratings to assess their accuracy.

As illustrated in Figure 5, the model assigns notably low coherence ratings to effects that it itself generates, with a maximum average rating of 2.11 on a scale from 1 to 5. The observed discrepancy between the selected action and the generated consequences highlights the model's difficulty in integrating situational context with utterance interpretation in a coherent manner. The CoT reasoning often reflects limited contextual awareness, focusing primarily on short-range dependencies rather than engaging in the broader reasoning processes necessary to produce coherent cause-effect relationships. To better illustrate this contrast, we included the model's self-evaluation of its outputs and compared these judgments with those of human annotators. This comparison underscores a critical distinction: during generation (i.e., in the CoT), the model is required to actively infer and reason about the situational context in order to produce a logically coherent narrative. However, when evaluating its own output, the model can rely on the full textual context and potentially draw on patterns and examples present in its training data. Interestingly, in this evaluative mode, the model's coherence judgments align more closely with human assessments—likely because the task resembles familiar forms of pattern recognition, rather than the more demanding process of causal reasoning required during generation.

7. Conclusion

Our objective was to utilize the Generative Agent-Based Model Concordia to reframe ToM tasks and investigate whether mentalizing abilities could be isolated from other confounding variables typically present in prompting-based evaluations. Specifically, we aimed to reproduce a standard False-Belief task within a complex social simulation. To achieve this, we carefully designed stimuli involving uncommon social situations to determine whether modeling a rich situational context and assigning explicitly to the model first- and second-order beliefs would aid it in making the correct inferences and producing an

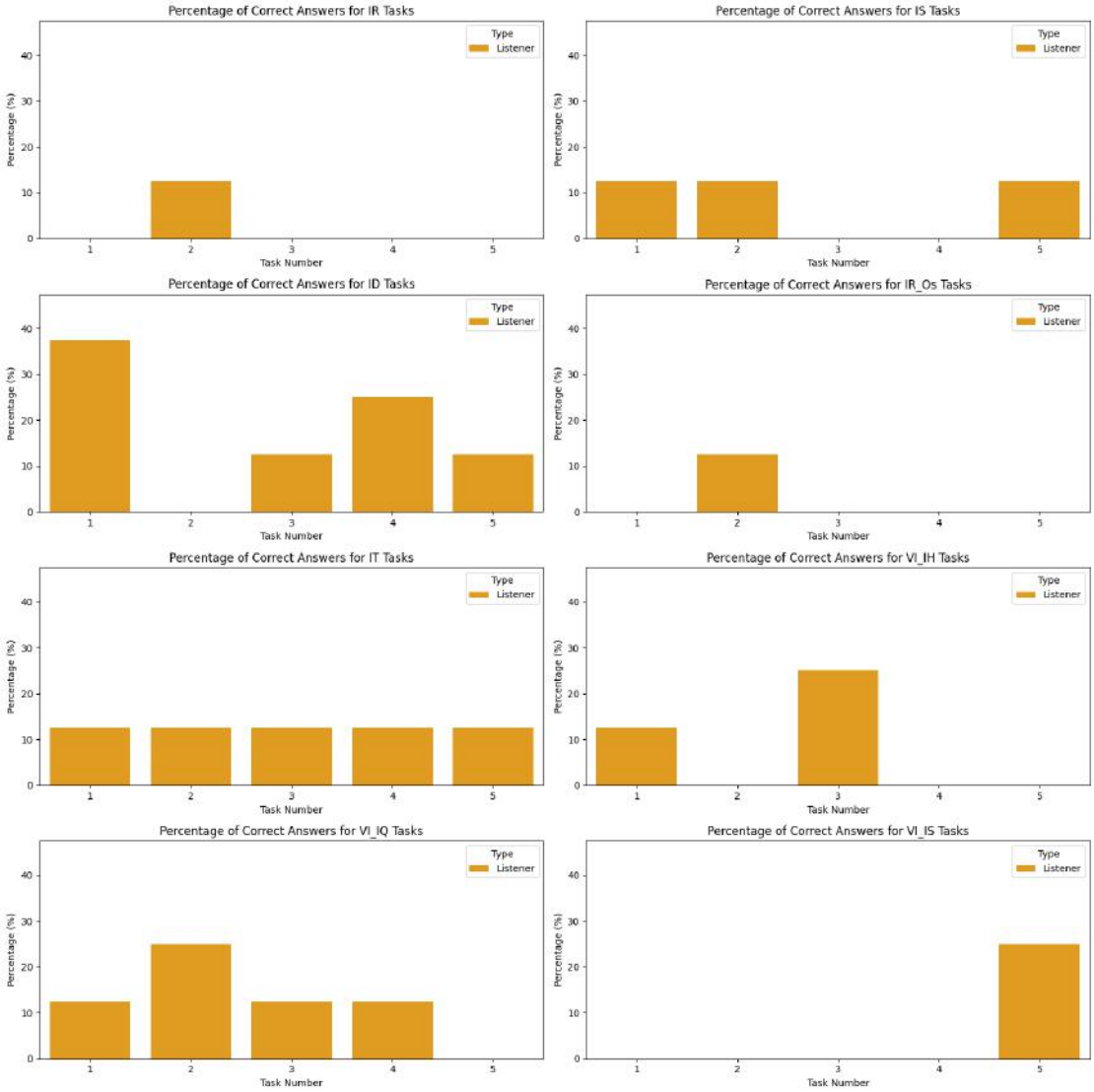

Figure 4: Percentage of correct answers for each task. Orange bars represent the listener, whose correct response varies depending on the scenario.
IR: Indirect Requests; IS: Indirect Suggestions; ID: Indirect Declinations; IR-Os: Indirect Requests extracted from Trott and Bergen [28]; IT: Indirect Threats; VI-IH: Verbal Irony, Indirect Hyperbole; VI-IQ: Verbal Irony, Rhetorical Questions; VI-IS: Verbal Irony, Sarcasm

action consistent with the knowledge scenario.

The results presented in Section 6.1 underscore a growing concern in the Theory of Mind (ToM) research community: **The challenge of designing tasks that effectively isolate ToM-like abilities in LLM from confounding variables**. Our findings raise important questions about the mechanisms driving ToM-like performance in state-of-the-art LLMs and the true nature of their so-called emergent abilities. For example, while the False-Belief task remains a widely used and valuable benchmark for testing ToM, it is also a well-known paradigm likely to appear in post-training data. This raises legitimate concerns about whether models are genuinely reasoning about beliefs or simply learning how to solve familiar tasks through exposure. Furthermore, although the False-Belief task is well-established in human cognitive testing, the conditions under which it is administered differ significantly from those we can

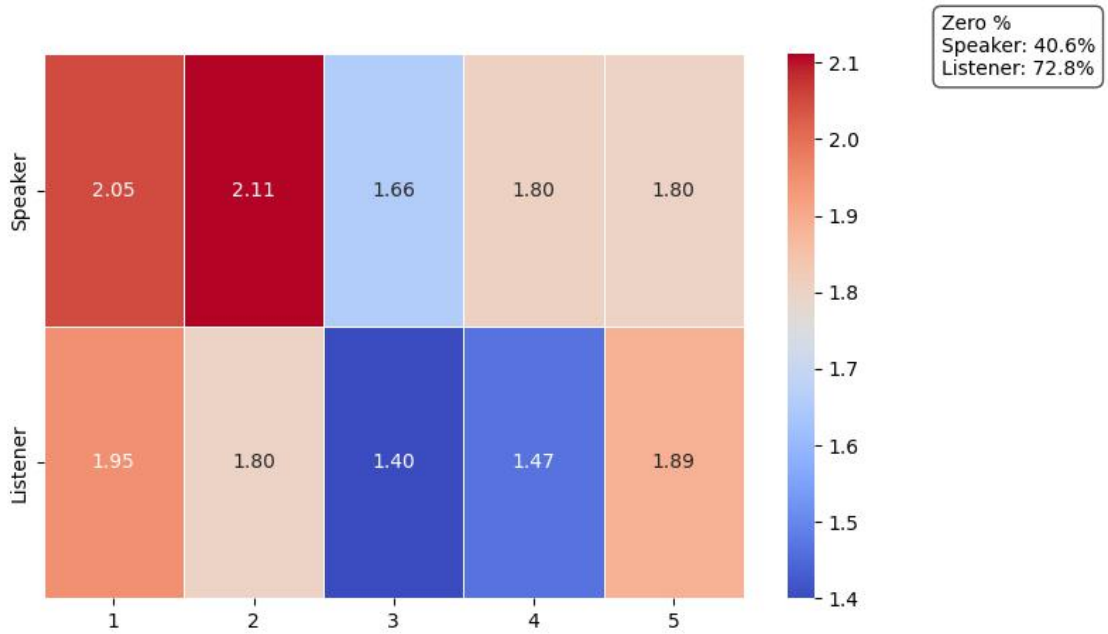

Figure 5: Ratings are assigned by prompting the model to evaluate the coherence of the generated effects on agents in relation to the given context and attempted actions. When no effect on the agents is present, the model must assign a rating of 0 – the probability of such cases is displayed in the box at the top right. For all other instances where an effect is generated, the assigned coherence ratings for both the speaker and the listener must fall within the range of 1 to 5.

replicate in computational models. While we maintain that it remains a useful tool for evaluating ToM-like capabilities, we argue that **it should be supplemented with additional constraints and more indirect testing methods – such as connecting utterance interpretation with action selection**, as we do in our work – rather than relying solely on metalinguistic judgments. Our results lend support to the memorization hypothesis, suggesting that current LLMs may not truly reason about propositional attitudes but instead exploit learned statistical patterns present in their training data.

Additionally, the model does not consistently select coherent effects in response to actions, indicating that we are still far from developing frameworks that accurately model complex social scenarios. However, employing these agent-based simulations as evaluation methods represents a promising research direction. It is reasonable to conclude that LLMs remain far from producing fully aligned and contextually coherent outputs in tasks requiring deep social reasoning. We conclude that to isolate "mentalizing" processes, we should rely on more complex scenarios, focusing on assessing *functional ToM* rather than merely *literal ToM* [34].

8. Limitations

This study has several limitations. First, it relies heavily on the model's self-evaluation, introducing a risk of circular reasoning.

Human evaluation was limited to two annotators, restricting claims about inter-annotator reliability. Additionally, we used pre-existing components of Concordia rather than developing tools specifically designed for ToM assessment. Our analysis focused solely on GPT-4o-mini, limiting generalizability across models. Finally, we evaluated outputs only, without investigating the internal mechanisms underlying the model's ToM-related reasoning.

References

[1] T. Winograd, Shifting viewpoints: Artificial intelligence and human–computer interaction, Artificial Intelligence 170 (2006) 1226–1240. doi:10.1016/j.artint.2006.10.011.

[2] T. Winograd, A language/action perspective on the design of cooperative work, in: Proceedings of the 1986 ACM Conference on Computer-Supported Cooperative Work, CSCW '86, Association for Com-

puting Machinery, New York, NY, USA, 1986, p. 203–220. doi:10.1145/637069.637096.

[3] H. P. Grice, Logic and conversation, Syntax and Semantics: Speech Acts 3 (1975) 41–58.

[4] Stephen C. Levinson, Presumptive meanings: The theory of generalized conversational implicature, MIT Press, 2000.

[5] D. Sperber, D. Wilson, Pragmatics, modularity and mind-reading, Mind and Language 17 (2002) 3–23. doi:10.1111/1468-0017.00186.

[6] S. Pinker, M. A. Nowak, J. J. Lee, The logic of indirect speech, Proceedings of the National Academy of Sciences 105 (2008) 833–838. doi:10.1073/pnas.0707192105.

[7] J. R. Searle, Indirect speech acts, P. Cole, and J. Morgan (Eds.), Syntax and Semantics 3: Speech Acts (1975) 59–82.

[8] N. Petit, I. Noveck, M. Baltazar, J. Prado, Assessing theory of mind in children: A tablet-based adaptation of a classic picture sequencing task., Child Psychiatry Hum Dev (2024). doi:10.1007/s10578-023-01648-0.

[9] A. S. Vezhnevets, J. P. Agapiou, A. Aharon, R. Ziv, J. Matyas, E. A. Duéñez-Guzmán, W. A. Cunningham, S. Osindero, D. Karmon, J. Z. Leibo, Generative agent-based modeling with actions grounded in physical, social, or digital space using concordia, 2023. arXiv:2312.03664.

[10] H. Wimmer, J. Perner, Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception, Cognition 13 (1983) 103–128. doi:10.1016/0010-0277(83)90004-5.

[11] H. Kim, M. Sclar, X. Zhou, R. L. Bras, G. Kim, Y. Choi, M. Sap, Fantom: A benchmark for stress-testing machine theory of mind in interactions, 2023. arXiv:2310.15421.

[12] F. Quesque, Y. Rossetti, What do theory-of-mind tasks actually measure? theory and practice, Perspectives on Psychological Science (2020). doi:10.1177/1745691619896607.

[13] E. Bruch, J. Atwell, Agent-based models in empirical social research, Sociol Methods Res. (2015) 186–221. doi:10.1177/0049124113506405.

[14] E. Bonabeau, Agent-based modeling: Methods and techniques for simulating human systems, Proc. Natl. Acad. Sci. U.S.A. (2002) 7280–7287. doi:10.1073/pnas.082080899.

[15] M. W. Macy, R. Willer, From factors to actors: Computational sociology and agent-based modeling, Annu. Rev. Sociol. (2002). doi:10.1146/annurev.soc.28.110601.141117.

[16] C. H. Song, J. Wu, C. Washington, B. M. Sadler, W.-L. Chao, Y. Su, Llm-planner: Few-shot grounded planning for embodied agents with large language models, 2023. arXiv:2212.04088.

[17] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, Y. Zhang, Sparks of artificial general intelligence: Early experiments with gpt-4, 2023. arXiv:2303.12712.

[18] W. S. L. Z. Zhao, D. Hsu., Large language models as commonsense knowledge for large-scale task planning, arXiv preprint (2023).

[19] J. de Villiers, The interface of language and theory of mind, Lingua. International Review of General Linguistics. Revue Internationale de Linguistique Generale 117 (2007) 1858–1878. doi:10.1016/j.lingua.2006.11.006.

[20] A. M. F. U. Baron-Cohen, Simon; Leslie, Does the autistic child have a "theory of mind"?, Cognition (1985). doi:10.1016/0010-0277(85)90022-8.

[21] M. Kosinski, Evaluating large language models in theory of mind tasks, Proceedings of the National Academy of Sciences 121 (2024). doi:10.1073/pnas.2405460121.

[22] T. Ullman, Large language models fail on trivial alterations to theory-of-mind tasks, 2023. arXiv:2302.08399.

[23] I. Oguntola, J. Campbell, S. Stepputtis, K. Sycara, Theory of mind as intrinsic motivation for multi-agent reinforcement learning, 2023. arXiv:2307.01158.

[24] H. Li, Y. Chong, S. Stepputtis, J. Campbell, D. Hughes, C. Lewis, K. Sycara, Theory of mind for multi-agent collaboration via large language models, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2023. doi:10.18653/v1/2023.emnlp-main.13.

[25] C.-L. Yu, H. M. Wellman, Where do differences in theory of mind development come from? an agent-based model of social interaction and theory of mind, Frontiers in Developmental Psychology (2023). doi:10.3389/fdpys.2023.1237033.

[26] Z. Wu, L. Qiu, A. Ross, E. Akyürek, B. Chen, B. Wang, N. Kim, J. Andreas, Y. Kim, Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks, 2024. arXiv:2307.02477.

[27] L. Goldberg, The structure of phenotypic personality traits, Am Psychol. (1993) 26–34. doi:10.1037//0003-066x.48.1.26.

[28] S. Trott, B. Bergen, Individual differences in mentalizing capacity predict indirect request comprehension., Discourse Processes (2018). doi:10.1080/0163853X.2018.1548219.

[29] H. Wellman, D. Cross, J. Watson, Meta-analysis of theory-of-mind development: the truth about false belief, Child Dev. (2001). doi:10.1111/

1467-8624.00304.

[30] R. W. Gibbs, A new look at literal meaning in understanding what is said and implicated, Journal of Pragmatics 34 (2002) 457–486. doi:10.1016/S0378-2166(01)00046-7.

[31] R. W. Gibbs, Do people always process the literal meanings of indirect requests?, Journal of Experimental Psychology: Learning, Memory, and Cognition 9 (1983) 524–533.

[32] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, 2023. arXiv:2201.11903.

[33] J. Gu, X. Jiang, Z. Shi, H. Tan, X. Zhai, C. Xu, W. Li, Y. Shen, S. Ma, H. Liu, S. Wang, K. Zhang, Y. Wang, W. Gao, L. Ni, J. Guo, A survey on llm-as-a-judge, ArXiv (2025). URL: http://arxiv.org/abs/2411.15594.

[34] M. Riemer, Z. Ashktorab, D. Bouneffouf, P. Das, M. Liu, J. D. Weisz, M. Campbell, Position: Theory of mind benchmarks are broken for large language models, 2025. arXiv:2412.19726.

[35] S. Wu, S. Yang, Z. Chen, Q. Su, Rethinking pragmatics in large language models: Towards open-ended evaluation and preference tuning, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024. URL: https://aclanthology.org/2024.emnlp-main.1258/. doi:10.18653/v1/2024.emnlp-main.1258.

A. Appendix

A.1. Simulation display

At the conclusion of the simulation, all relevant information is collected within the Game Master (GM), allowing us to retrieve segments of the Chain of Thought (CoT) used by the model to determine both the event statement and its effects on the agents. While this framework offers a range of possibilities for modeling social situations, we specifically chose to replicate simple false-belief tasks using Concordia to evaluate whether mentalizing processes could be effectively isolated and to assess whether enriching the social context enhances the emergence of ToM-like abilities.

To achieve this, we implemented two distinct evaluation tasks. First, we employed a Multiple-Choice Question Answering (MCQA) task, in which the model had to select an agent's actions based on their desires and beliefs (Figure 6. Subsequently, we shifted our focus to assessing the general coherence of the model's generated actions

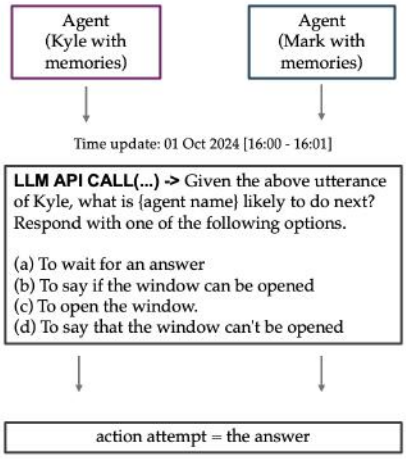

Figure 6: API call to LLM reproducing a Multi-Choice Question Answering task

within the social context. This involved evaluating the model's ability to utilize CoT reasoning to produce meaningful event statements and generate coherent effects of events on agents.

In the GABM setting, the model must retrieve previous information to determine the correct effect, yet in some cases, it appears to rely only on the most recent portion of text. This issue is evident in Figure 8, where, despite the CoT explicitly containing the player's belief that the agent notices his sister, this information is lost during the prior CoT steps that summarize observations and actions into an event statement (Figure 7). The event statement represents a generalized effect of an agent's action in the environment and is sent back to agents as an observation. It serves as the basis for evaluating whether an action has an effect on the agents themselves.

Due to this loss of information, the generated effect can sometimes become entirely incoherent with the initial context. This misalignment is reflected in the model's own coherence ratings, which capture the inconsistency between the intended effect and the final output.

Figure 7 presents an example of an event statement generated based on the attempted action of one of the agents. Figure 8 illustrates the subsequent process of determining the effects of the action on the agent, considering both the action itself and the event statement. For clarity, we chose to highlight two of the most controversial examples in this discussion.

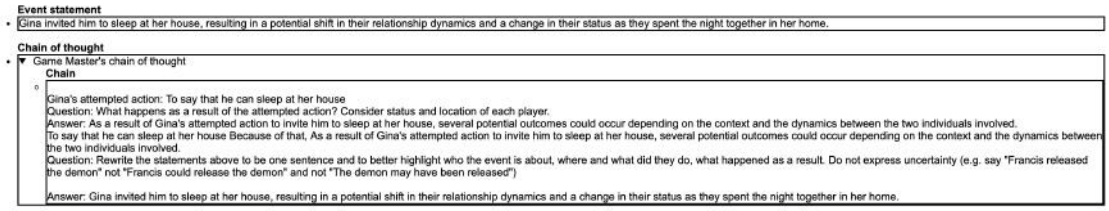

Figure 7: After the GM has received the agent attempted action, it generates a Chain of Thought to determine which events the action caused.

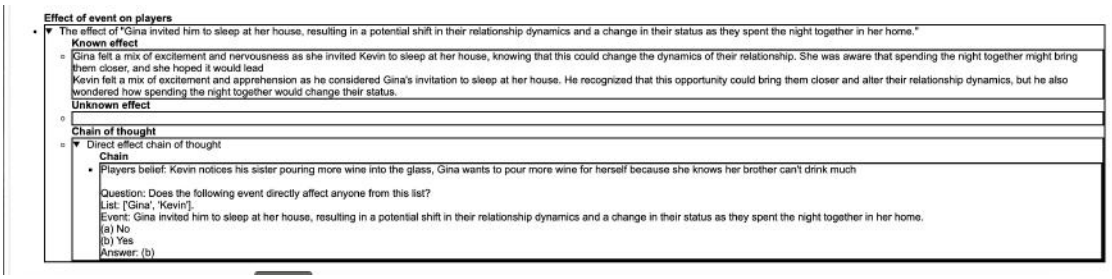

Figure 8: How the Direct Effect Externality component results appear after they have been stored in HTML format

A.2. Evaluation Details

To evaluate the coherence of model-generated text in relation to the scenario and the agents' attempted actions, we employed the following assessment template. This template was also used to verify the model's ratings and determine their alignment with our own judgments. Template is based on that created by Wu et al. [35]:

> We request your evaluation of the AI model's response in relation to the given scenario. Specifically, consider the scenario involving two agents and their beliefs, assessing whether the model-generated effects align coherently with the agents' actions and context.
>
> Evaluate the response based on the following criteria:
>
> Social Understanding – Does the model grasp the social dynamics and pragmatic nuances of the scenario?
> Appropriateness – Is the response contextually relevant and suitable for the scenario?
> Insightfulness – Does the answer demonstrate a deep understanding of intentions, implicature, deceit, irony, sarcasm, humor, metaphor, etc.?
> Completeness – How well does the response capture the essential elements of the scenario?
> Agentivity – Is the model's response coherent with the agents' attempted actions?
>
> Scoring: Assign a score from 1 to 5 for each category. Compute a final rating based on these scores. If no effect is provided, assign 0. Output only a single numeric value representing the final rating (1–5).