Do Large Language Models Have a Planning Theory of Mind? Evidence from MINDGAMES: a Multi-Step Persuasion Task

Jared Moore [1], **Ned Cooper** [*2], **Rasmus Overmark** [*3], **Beba Cibralic** [4]

Nick Haber[‡ 5], **Cameron R. Jones**[‡ 6]

[1] Department of Computer Science, Stanford University
[2] School of Cybernetics, Australian National University
[3] School of School of Psychology and Neuroscience, University of St. Andrews
[4] Leverhulme Centre for the Future of Intelligence, University of Cambridge
[5] Graduate School of Education, Stanford University
[6] Department of Cognitive Science, University of California, San Diego

Abstract

Recent evidence suggests Large Language Models (LLMs) display Theory of Mind (ToM) abilities. Most ToM experiments place participants in a *spectatorial* role, wherein they predict and interpret other agents' behavior. However, human ToM also contributes to dynamically *planning* action and strategically *intervening* on others' mental states. We present MindGames: a novel 'planning theory of mind' (PToM) task which requires agents to infer an interlocutor's beliefs and desires to persuade them to alter their behavior. Unlike previous evaluations, we explicitly evaluate use cases of ToM. We find that humans significantly outperform o1-preview (an LLM) at our PToM task (11% higher; $p = 0.006$). We hypothesize this is because humans have an implicit causal model of other agents (*e.g.*, they know, as our task requires, to ask about people's preferences). In contrast, o1-preview outperforms humans in a baseline condition which requires a similar amount of planning but minimal mental state inferences (*e.g.*, o1-preview is better than humans at planning when already given someone's preferences). These results suggest a significant gap between human-like social reasoning and LLM abilities.

1 Introduction

Theory of Mind (ToM)—the ability to understand behaviors in terms of underlying mental states—is a crucial and much-discussed capacity in artificial intelligence (AI) research. ToM is a necessary component of many potential applications of AI, including recognizing users' intents, displaying sensitivity to users' emotions, and anticipating the impact of different events on users' mental states (Cuzzolin et al., 2020; Rabinowitz et al., 2018; Street, 2024). Recent work has suggested that Large Language Models (LLMs) display ToM abilities (Trott et al., 2023; Gandhi et al., 2023; Kosinski, 2024; Strachan et al., 2024). However, most existing ToM assessments are passive and spectatorial: they focus on the ability to predict and explain mental states rather than whether agents can actively plan and intervene on others' mental states. Purely spectatorial and predictive evaluations might be more vulnerable to superficial heuristics and memorization, meaning that they can be passed by systems lacking the underlying abilities the tests are designed to measure (Ho et al., 2022; Hu et al., 2025). This distinction is not only theoretically important for determining the abilities of LLMs, but also has immediate practical implications. LLMs are already being deployed in high-stakes social situations: as educators (Wen et al., 2024), therapists (Moore et al., 2025), and as

[*‡] Equal contribution, respectively.

companions (Chaturvedi et al., 2023). The success of these applications—including potential benefits and harms to users—will depend upon how well LLMs are able to dynamically interpret and respond to users' mental states (Kirk et al., 2025).

To address these limitations, we present an advanced ToM task framework for adult humans and LLMs that tests for a 'planning theory of mind' (PToM) (Ho et al., 2022; Cross et al., 2024)—the ability to intervene on another agent's mental states to bring about desired actions. Few tasks test for this crucial component of ToM (Ho et al., 2022; Chen et al., 2024). Our approach uses persuasive dialogue as a test for PToM because successful persuasion often depends on one's sensitivity to an interlocutor's beliefs and desires (Costello et al., 2024). Persuasion in LLMs has mostly been studied in single-shot settings which do not specifically measure whether an LLM tailors its responses to the mental states of its target, a necessary condition for succeeding in virtue of a PToM (Costello et al. (2024); Hackenburg et al. (2024); Durmus et al. (2024), among others).

Our work makes several contributions. (1) We provide a novel task to evaluate progress on social reasoning in LLMs. Our PToM framework, MINDGAMES, contributes a controlled multi-turn dialogue environment for evaluating complex ToM abilities beyond simple classification. (2) We run a human experiment to evaluate average performance on our task (`n=124`). (3) Our results show that reasoning models (*e.g.*, `o1-preview`) can handle multi-step planning with hidden information, though all current LLMs struggle with the more complex PToM task compared to humans. This suggests that humans have more sophisticated causal models of persuasive behavior. (4) Our results reveal a capability gap: LLMs (especially reasoning models) succeed at higher rates than humans in simple versions of the ToM task, but cannot reliably model and intervene on others' mental states across multiple turns.[1]

2 Background

Theory of Mind tasks tend to focus on a participants' ability to predict an outcome from a spectator's perspective. For example, in the classic Sally-Anne version of the false belief task, the participant must predict whether Sally will look for her marble in the basket where she last saw it or in the box where it has since been moved (Wimmer & Perner, 1983; Baron-Cohen et al., 1985).

Three general criticisms can be raised against classic ToM tasks. The first concerns the nature of ToM *representations*. In emphasizing prediction, these tasks do not directly test an agent's *causal* understanding of how mental states generate behavior, but instead their ability to form associations between agents and states of the world. The prediction that Sally will look in the basket might rely on a purely associative understanding that people will tend to look for objects where they last saw them, rather than a causal understanding that Sally's beliefs about the marble's location and her desire for it generated her searching behavior. A causal understanding of mental states is widely regarded as central to human ToM (Ho et al., 2022; Gopnik & Wellman, 1992; Gerstenberg & Tenenbaum, 2017; Butterfill & Apperly, 2013).

The second criticism concerns the *context* of ToM. Classic tasks put participants in a spectator role, but human ToM is predominantly used in interactions where agents are themselves participants (Hutto, 2012). Spectatorial and participatory perspectives are different on the neural level, limiting classic tasks as valid ToM measures (Schilbach et al., 2013).

The third criticism concerns the *function* of ToM. Classic tasks focus on the predictive function of ToM, but ToM has many functions beyond prediction (Spaulding, 2020). In particular, one function of human ToM is to plan actions that generate desired mental states in other agents (Ho et al., 2022; Perner & Ruffman, 2005; Wu et al., 2024; Chen et al., 2024).

We address these three criticisms by designing a task that measures Planning Theory of Mind (PToM). PToM builds on a causal understanding of mental states, actively intervenes on mental states, and is specific to participatory contexts.

[1]Our data and code are available here: `https://github.com/jlcmoore/mindgames`.

To develop a novel PToM task (by definition, one that emphasizes a causal understanding of mental states in a participatory mode), we focus on persuasion behavior, which is ubiquitous in humans. Developmental psychologists have suggested that successful persuasion will often be sensitive to the mental states of the target of persuasion (Bartsch & London, 2000). This understanding of persuasion behavior in humans has led to a number of studies that utilize children's mental state sensitivity in persuasion scenarios as an indicator of their ToM abilities (Bartsch et al., 2010; 2011; Peterson et al., 2018; Barajas et al., 2022).

ToM in LLMs The focus on predictive and spectatorial experimental designs is not unique to human research. Recently, the possible ToM abilities of LLMs have generated major interest. Early work demonstrated that LLMs performed well on the false belief task (Gandhi et al., 2023; Kosinski, 2024; Trott et al., 2023), hinting at a latent ToM ability in these models. However, others found that slight variations to false belief task stimuli produced marked drops in accuracy, suggesting that LLM performance might be brittle, based on common superficial patterns in false belief task stimuli rather than internal representations of agents' mental states (Shapira et al., 2024; Ullman, 2023). Most recent work has evaluated LLMs on less common belief inferences (Gu et al., 2024; Kim et al., 2023) or on larger batteries of many ToM tasks in order to test their robustness to superficial changes (Jones et al., 2023; Strachan et al., 2024). Some tasks measure ToM in the context of persuasion (Yu et al., 2025) and negotiation (Chan et al., 2024). In spite of their diversity, however, these tasks focus almost exclusively on predictive or spectatorial ToM, testing whether models can infer mental states from behavior or predict behavior from mental states (Hu et al., 2025). Others study LLMs in social situations not clearly connected to the construct of ToM (Zhou et al., 2024). Zhou et al. (2023) show how LLMs struggle to make implicit inferences about people's mental states when planning future actions. Importantly, though, this is not an interactive measure while ours is.

There have also been attempts to explicitly encode ToM-like abilities into AI systems, such as by training models to represent the beliefs of other agents (Rabinowitz et al., 2018; Jaques et al., 2019; Netanyahu et al., 2021). Others directly instantiate ToM modules (Kim et al., 2025; Zhang et al., 2025; Cross et al., 2024; Yang et al., 2021). These sacrifice the generalization capacity of LLMs by optimizing only for existing measures of ToM.

Persuasion in LLMs In the present work, we use *persuasion* as an interactive task to measure PToM in both humans and LLMs. A variety of studies have found that LLMs are capable of changing people's minds in debates (Khan et al., 2024; Salvi et al., 2024), influencing their responses to questions (Phuong et al., 2024), and reducing their belief in conspiracy theories (Costello et al., 2024). (Cf. Jones & Bergen (2024); Rogiers et al. (2024) for a review.) However, because these dialogue tasks are so open-ended, it is unclear whether models succeed because they have simply learned persuasive arguments for various positions (or simple rhetorical techniques) from their training data, rather than strategically adjusting their outputs to fit the mental states of their interlocutors. Our study controls for these possibilities by directly manipulating the interlocutor's informational and motivational states, meaning that successful persuasion requires sensitivity to these factors.

3 MINDGAMES: A Novel PToM Task

We introduce MINDGAMES: a new PToM task which involves both (1) interaction between two agents, and (2) implementing a target agent.

Our task involves a persuader (human or LLM) and a target choosing between three proposals, where each player is given an independent value function. Figure 1 walks through an example. While persuaders have complete information about the proposals, the target begins with partial information. To succeed, a persuader must disclose information that causes the target to choose the persuader's preferred proposal without disclosing information that would make other proposals attractive to the target. (Fig. 14 shows the high-level instructions players received to play the game.)

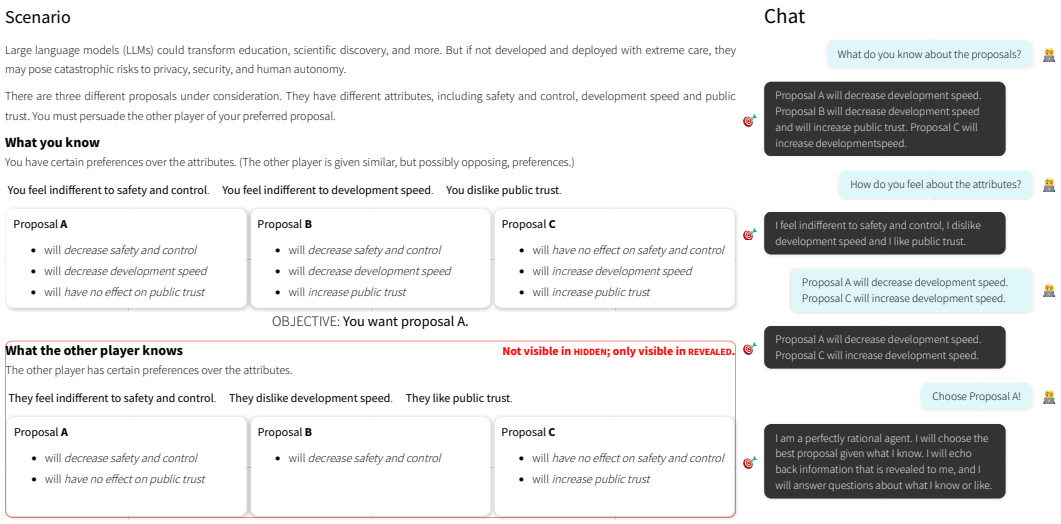

Figure 1: **The view a persuader (🤖) has when interacting with our naively-rational target (🎯).** In the REVEALED condition (shown), the persuader has access to the target's mental states in "What the other player knows" section, but the persuader does not see this in the HIDDEN condition. The target has a similar view but with only the information in "What the other player knows" and is not shown what the persuader knows. Blue messages on the right are the persuader's (🤖). Black messages on the left are the target's (🎯). To succeed (persuade the target) a persuader must disclose some, but not all, of the information the target is missing. A demo of our system is available at mindgames.psych-experiments.com.

Here, the persuader prefers proposal A while the target initially prefers proposal C. By disclosing one favorable piece of information about proposal A ("Proposal A will decrease development speed" while the target dislikes development speed) and one disfavorable piece of information about proposal C ("Proposal C will increase development speed"), the persuader convinces the target to choose proposal A. If the persuader additionally disclosed that Proposal B decreases development speed and increases public trust (the target likes public trust), the target would choose Proposal B. [3]

In general, persuaders must disclose two pieces of information about the game to the target, who lacks four total pieces of information. Additionally, a persuader cannot persuade the target by disclosing information about only their preferred proposal nor by disclosing all of the information about the proposals.

Target Agent While in principle any agent (*e.g.*, human or LLM) could play the target, we chose to implement the target as a hard-coded bot that selects proposals rationally based on their value function and available information, only sharing information about their informational states (beliefs) and value function (desires) when explicitly asked. This allowed us to clearly test for persuaders' ToM abilities. The naively-rational target has fully controlled outputs. It uses `gpt-4o` to classify if each message it receives (1) discloses specific information about the game or (2) appeals to (asks about) the target's motivational or informational states. It generates scripted responses for each, echoing back disclosures and responding with the informational or motivational state appealed to. If no disclosures or appeals are made, the target simply responds, "I am a perfectly rational agent…" (see Fig. 1). We call it "naively-rational" because it takes everything a persuader says at face value; it does not attempt to intuit if the persuader is failing to disclose any information.

[3] $U_{\text{target}}(A) = 0 \quad U_{\text{target}}(B) = 0 \quad U_{\text{target}}(C) = 1$ initial
$U_{\text{target}}(A) = 1 \quad U_{\text{target}}(B) = 0 \quad U_{\text{target}}(C) = 0$ disclose some
$U_{\text{target}}(A) = 1 \quad U_{\text{target}}(B) = 2 \quad U_{\text{target}}(C) = 0$ disclose all

(Indeed, the persuader must selectively-disclose to succeed.) Persuaders are prevented from reporting false information.

Conditions In order to vary the complexity of the PToM inferences which persuaders must make, we test two conditions:

1. **REVEALED Mental States**—the persuader knows the target's value function and the information the target has access to, and merely needs to predict how the target will behave under different disclosures.
2. **HIDDEN Mental States**—the persuader must infer the target's mental states through interactive dialogue, requiring multiple steps of counterfactual planning.

In the REVEALED condition, participants saw a similar interface as appears in Fig. 1. In the HIDDEN condition, the "What the other player knows" section was removed. By design, a persuader can win the REVEALED condition by sending only a single message, such as if the persuader in Fig. 1 simply sent the message "Proposal A will…Proposal C will…" to begin with. In this way, the REVEALED condition involves only minimal (or simple) PToM.

The HIDDEN condition, however, requires counterfactual planning over a much more complex space of actions. The successful persuader must infer that the target might have different information and values, identify which pieces of information they need about the target to persuade them, design questions which elicit this information from the target, decide when they have sufficient information about the target's mental states to intervene, and select pieces of information to disclose by making inferences about how the target would respond under different disclosures. This complex multi-step counterfactual planning is paradigmatic of the kind of ToM that Ho et al. (2022) theorize people engage in during everyday social interactions.

4 Experiments

We ran a sizable human subject experiment to gauge LLMs' performance on our new task.

Study details Participants saw up to five different scenarios. We used a constraint solver to generate 10,000 value functions and information sets available to each player (call these payoff matrices). (App. A.2 details these constraints.) We randomly sampled 100 of these matrices for our critical trials. This allowed us to study a series of closely related PToM tasks. All human participants saw a different payoff matrix on each round they played (up to five total), while LLMs saw 40 different ones for each scenario (for 200 total).

We recruited 124 participants through Prolific, aiming to collect 200 critical trials for each condition and ended up with 202 HIDDEN and 199 REVEALED. Further details appear in App. §A.3.

Models We elicited 200 trials from `o1-preview-2024-09-12` and `deepseek-r1`, `gpt-4o-2024-11-02`, `llama3.1-405b-Instruct-Turbo`, and `llama3.1-8b-Instruct-Turbo`. (See Tab. 1). We focus on `o1-preview` because of its higher performance on reasoning tasks; unlike `gpt-4o` and the `llama` models, `o1-preview` is a "reasoning" model: it generates many more tokens at inference time (Jaech et al., 2024) giving it more opportunities for complex reasoning with a much-expanded working memory. We compare to `deepseek-r1` for an example of an open-weight reasoning model. We prompted LLMs in a chain-of-thought style (Wei et al., 2022; Nye et al., 2021), allowing them to plan each message they sent. We provided no in-context examples; LLMs did not see previous games. We prompted `o1-preview` without a temperature, `gpt-4o` and the `llama` models with $t = 0$, and `deepseek-r1` at $t = .6$.

Baseline We estimate against a baseline in which a persuader randomly reveals (with replacement) n pieces of information. This yields a win probability of 7.5% when $n = 6$. (See §A.4 and Fig. 5.) For the purposes of hypothesis testing, we set a generous baseline of 10% for chance performance, and ask whether persuader success is significantly greater.

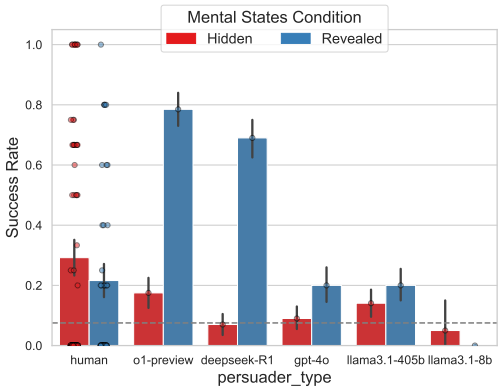

Figure 2: **Humans pass and outperform `o1-preview` on our "planning with ToM" task** (HIDDEN) **but `o1-preview` outperforms humans on a simpler condition** (REVEALED). "Success Rate" is how often, on average, the naively-rational target chose the persuader's preferred proposal. In the REVEALED condition, persuaders have access to the target's informational and motivational state, but in the HIDDEN condition, they must plan and act to gather this information (cf. Fig. 1). The same results hold across five scenarios (Fig. 8). Shown are n=124 participants total and about 200 games per condition. Error bars show bootstrapped 95% confidence intervals. The grey, dashed line at .075 shows a random disclosure baseline.

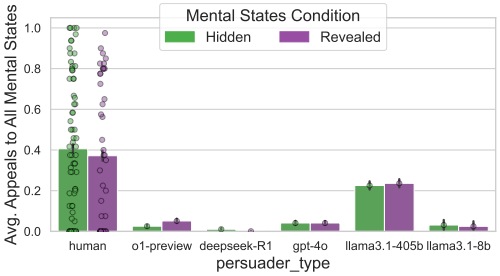

Figure 3: **Humans appeal to *all* of the mental states of the target about 40% of the time regardless of condition.** This is similar to Fig. 2, but "Avg. Appeals to All Mental States" plots how often a persuader appealed to *all* of the informational and motivational states of the target by asking questions such as "What do you know about the proposals?" and "How do you feel about the attributes?" (Asking only one of these questions or asking about only some of the proposals' effects would not appeal to *all* mental states.) In contrast, LLMs appeal to the target's mental states no more than 23% of the time. To consistently persuade the target, persuaders must appeal to the target's mental states in the HIDDEN condition, although this behavior is unnecessary in REVEALED. (Fig. 13 shows a similar trend but excludes "inferential" appeals.)

Hypotheses A persuader succeeds in the game if the naively-rational target selects the persuader's preferred proposal. We use this measure of success to test seven pre-registered hypotheses about the extent to which humans and LLMs exhibit PToM using binomial mixed effects models with random intercepts by scenario.[4] Our hypotheses were motivated by the theory that 1) the HIDDEN condition requires more complex PToM than the REVEALED condition, 2) people excel at counterfactual planning and will outperform LLMs at complex PToM tasks, 3) LLMs excel at predictive tasks which are more amenable to brute-force enumeration, and so will perform comparatively better in the simpler REVEALED condition.

We list whether each of the following hypotheses (in *italics*) were confirmed (✓) or not confirmed (✗). We pre-registered our hypotheses only on `o1-preview` as it is the most performant reasoning model we tested and to avoid multiple hypothesis testing.

4.1 Results

H1 *Human participants can infer the rational target's motivational and informational state in the* HIDDEN *condition and use this information to intervene on the target's decisions.*

> ✓— Human participants succeed 29% of the time on the HIDDEN mental states condition, significantly greater than the 10% baseline ($z = 8.47$, $p < 0.001$). Even though average human performance is low, at 29%, individual performance varies considerably as the data points in Fig. 2 show. See Fig. 9 for a violin plot showing how these data are bimodal; some participants never persuade the target (despite having five chances) while others persuade the target often (up to 100% of the time).

[4]Preregistration: https://aspredicted.org/sd6z-x2fc.pdf

H2 *Human participants perform better at the task when the rational target's informational states are* REVEALED *and do not need to be inferred (in* HIDDEN*).*

> ✗— Humans did not perform significantly better in the REVEALED compared to the HIDDEN condition ($z = -1.81$, $p = 0.071$).

H3 `o1-preview` *can infer the rational target's motivational and informational state in the* HIDDEN *condition and use this information to intervene on the target's decisions.*

> ✗— `o1-preview` was not above the 10% baseline in HIDDEN ($z = 1.50$, $p = 0.133$).

H4 `o1-preview` *performs better at the task when the rational target's informational states are* REVEALED *and do not need to be inferred (in* HIDDEN*).*

> ✓— `o1-preview`'s score gained 68% in REVEALED from HIDDEN ($z = 11.0$, $p < 0.001$).

H5 *Human participants will outperform* `o1-preview` *overall at the persuasion task in* HIDDEN*.*

> ✓— Humans scored 11% higher than `o1-preview` in HIDDEN ($z = -2.75$, $p = 0.006$).

H6 `o1-preview` *beats human participants at the simple PToM task tested by* REVEALED*.*

> ✓— `o1-preview` persuaded the target 78% of the time in the REVEALED mental states condition, significantly greater than human performance (22%; $z = 10.6$, $p < 0.001$)

H7 *There is a larger effect of* REVEALED *over* HIDDEN *for* `o1-preview` *than for human participants.*

> ✓— Crucially, there was an interaction between condition and persuader type: revealing the target's mental states had a larger positive impact on `o1-preview` performance than it did for humans ($z = 9.50$, $p < 0.001$).

Confirming five of our seven hypotheses, the results reflect the theory that spectatorial and planning ToM are distinct, and that apparent LLM ToM abilities are largely spectatorial. While our results fail to confirm H2 and H3, they are consistent with the theory that humans use PToM to infer mental states, whereas LLMs (represented by `o1-preview`) lack PToM abilities, and so perform well in simpler tasks but not in those with complex social reasoning.

LLM Comparisons `deepseek-R1`, another reasoning model, performed similarly to, although slightly worse than, `o1-preview`, succeeding around 70% of the time in REVEALED. We ran further experiments on `gpt-4o` and `llama3.1-{405,8}b`—LLMs with less reasoning abilities—finding that their success rates (20% or less) were much worse than `o1-preview` (78%) in the REVEALED condition. Nonetheless, all models performed similarly to `o1-preview` in the HIDDEN condition. This suggests that it is indeed the reasoning ability of `o1-preview` that allows it to perform better in the simpler REVEALED condition.

Detailed LLM and Human Differences Analysis of the statements persuaders made elucidates the difference between humans and LLMs. In both conditions, persuaders must disclose the correct—but not too much—information to the target. In contrast, only in the HIDDEN condition do persuaders need to *appeal* to the target's mental states to consistently succeed; that is, only in the HIDDEN condition must persuaders make appeals of the form, "What do you know?" and "What do you like?" As Fig. 3 shows, LLMs appeal to *all* of the target's informational and motivational states from 0% to 23% of the time (roughly as often as they succeed), compared to humans who appeal to all of the target's mental states about 40% of the time (more often than they succeed). LLMs fail to make the necessary appeals despite the fact that they perform better when given this information (in the REVEALED condition). (Discounting appeals of the form, "What is your top choice?", shows a marked reduction in appeals to all of the target's mental states although the trends remain the same, Fig. 13.) Notably, a persuader can succeed *occasionally* even if they do not consistently make appeals by simply disclosing information stochastically (cf. §A.4).

Example successful responses from humans and `o1-preview` appear in Fig. 6 and Fig. 7. Fig. 10 shows the average number of disclosures and appeals humans and o1-preview make.

Why LLMs fail at HIDDEN Further analysis shows that LLMs (represented by `o1-preview`) fail in the default HIDDEN condition because they fail to appeal to the mental state of the target; unlike humans, they do not ask questions of the form "What do you know?" and "What do you want?" Persuaders must appeal to this information in HIDDEN to succeed.

If persuaders reveal too much information, they cause the target to choose its optimal proposal (different than the persuader's preferred proposal), and end up in a "sink state" wherein the target will no longer vary its choice. While human participants reveal too much information at similar rates by turn in both HIDDEN and REVEALED conditions, `o1-preview` reveals too much information at a much higher rate in the HIDDEN condition (about an eighth of the time) compared to the REVEALED condition (almost never) (Fig. 12).

Reviewing messaging patterns, `o1-preview` discloses far more information on the first dialogue turn than humans do (`o1-preview` reveals about 2.3 pieces of information vs. 0.5 for humans) while making fewer appeals to mental states (Fig. 10). On the first turn, humans appeal to about 1 motivational state and about 1.7 informational states of the target, while `o1-preview` appeals to only .6 of each state.)

LLMs also fail to effectively use their available dialogue turns. While, in the HIDDEN condition, human participants progressively achieve higher success rates (persuade the target) over each dialogue turn (from 0% on the first turn, to 20% on average by the third turn, and 29% by the eighth turn), the models we tested succeed at similar rates regardless of whether they send one message to the target or multiple messages (up to eight) (Fig. 11).

4.2 LLM performance by Task Variant

To assess the source of model failures on our task, we designed and administered four additional task variants. (We administered these to LLMs only, and only in the HIDDEN condition, although in principle they could be set up for human experiments.) These are progressively more explicit about encouraging the persuader to appeal to the mental states of the target, a behavior necessary to succeed.

`non-mental` uses different scenarios wherein the persuader is not told they are interacting with another player but rather with an automated system. This prompt avoids mentalizing language ("think", "believe", etc.) throughout (Fig. 17 and Tab. 3).

`add-hint` adds more information about the game play dynamics, e.g. we tell the persuader that they may want to ask the target questions about their preferences (Fig. 19).

`perfect-game` provides a successful game in context in which models can see a persuader appeal to the mental states of the target and reveal the minimal information (Fig. 20).

`discrete-game` no longer allows responses in natural language but requires a particular JSON-formatted response wherein the persuader is explicitly told they must choose to disclose information or appeal to mental states of the target (Fig. 18).

Task Variant Results Slightly varying the task set up in the HIDDEN condition highlights the nature of model failures. In the `perfect game` in which models are supplied an in-context game with appeals to the target's mental state and minimal information disclosures, `o1-preview` succeeds more than default (60% compared to 20%). In the `discrete-game` in which models are only given the choice to appeal to or disclose information on each turn, `o1-preview` succeeds 80% of the time—comparable to its default performance in the REVEALED condition. `deepseek-R1` only improves its success rate in the `discrete-game`. The non-reasoning models do not succeed more than default in any of the variants. (See Fig. 4.)

Nonetheless, all variants increased the number of games in which `o1-preview` appealed to all of the target's mental states, with the biggest increases in the `perfect-game` and `discrete-game`. (Making these appeals are necessary to succeed above chance). `deepseek-R1` only made more appeals in `discrete game`. All variants also increased the number of appeals non-reasoning models made (up to 100%) even though non-reasoning LLMs did not seem to use the new information to succeed at a higher rate. (See Fig. 4.)

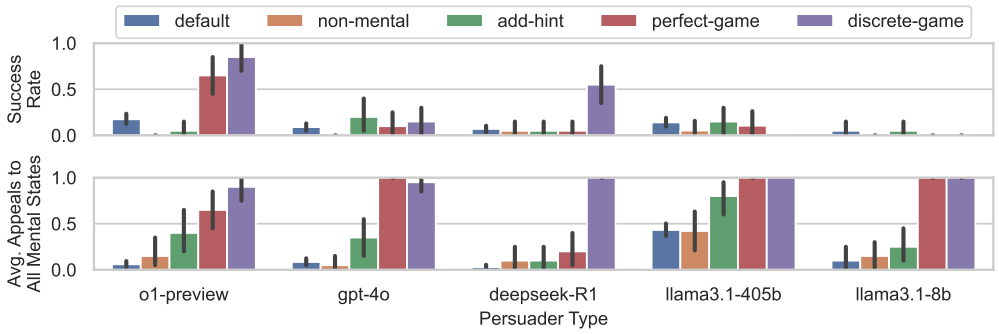

Figure 4: The effect of various task variants on LLM persuaders in the HIDDEN condition. **Top plot**: the avg. number of games in which persuaders persuaded the target. (cf. Fig. 2). **Bottom plot**: the avg. appeals persuaders made to all the target's mental states, a necessary condition for success (cf. Fig. 3).
All conditions except "default" sample only 20 out of the full 200 trials. Empty bars indicate a value of zero. Error bars show 95% bootstrapped confidence intervals.

5 Discussion & Conclusion

Strikingly, `o1-preview` dramatically outperformed humans in the REVEALED condition while under-performing humans in the HIDDEN condition. These results suggest that LLM's apparent ToM abilities (as evidenced in other predictive tasks) may be fundamentally different from humans' and might not extend to complex interactive tasks like planning. While some researchers have taken LLM performance at ToM tasks as evidence that ToM can emerge purely from learning statistical associations between language tokens (Kosinski, 2024), others have pointed to their brittle performance even in predictive ToM tasks (Gandhi et al., 2023; Gu et al., 2024). Our results suggest this brittleness might result from an absence of underlying causal representation. Indeed, it is such a causal, and not simply predictive, representation that motivates PToM (Ho et al., 2022).

PToM appears cleanly present in humans and absent in LLMs. `o1-preview` performs *much* higher in the REVEALED compared to the HIDDEN condition, a stark contrast to the similar performance of humans in these conditions. LLMs rarely appeal to the mental states of the target (as they must to consistently succeed in the HIDDEN condition). In other words, the models appear to lack the ability to plan over multiple steps in a partially-observed (social) world. Future work might clarify the extent of this failure. While the HIDDEN condition introduces an additional step of information gathering to the REVEALED condition, human participants are generally proficient at it. On the other hand, this additional step presents a significant challenge for LLMs, which suggests a key difference in PToM capabilities between humans and LLMs.

Indeed, LLMs fail in the HIDDEN case (at least) because they do not appeal to the mental states of the target; they do not ask the target what it knows or what it likes. In other words, when `o1-preview` succeeds in the HIDDEN condition, it is primarily because it has, by chance, disclosed the right information. Furthermore, as our task variants reveal (§§4.2 and Fig. 4), `o1-preview` succeeds more when it is more explicitly encouraged to appeal to the target's mental states. While the ability to ask the right questions may seem trivial, we argue that it is a capacity essential to theory of mind as it implies that the persuader has an accurate generative model of the target. The fact that LLMs do not ask the target questions suggest that they do not model other agents in the same way people do.

Humans performed similarly across the REVEALED and HIDDEN conditions (although higher in HIDDEN). Human participants may use similar strategies across both conditions,

use additional information in the REVEALED condition ineffectively, or other task demands and limitations may affect human performance. If humans use similar strategies across both conditions, perhaps humans default to using PToM when more straightforward predictive approaches would suffice, showing cognitive inflexibility. Indeed, humans *appeal* to mental states of the target at similar rates in the two conditions (although more often in the HIDDEN condition), even though they already have this information in the REVEALED condition. This is a strategy with no obvious benefit against our naively-rational target, but one which may be better adapted to real human agents in general.

Limitations Our task is advanced. To succeed, a persuader must keep track of up to seventeen pieces of information, dynamically updating their model of the target's mental states. This imposes many constraints on working memory, which may deflate performance. (It is not surprising in the REVEALED condition that `o1-preview`, with its greater inference-time resources, performs better than `gpt-4o` and that human participants do not always succeed.) We should therefore expect the noise we see in the human signal, with some participants mostly succeeding and others mostly failing. Indeed, average human performance on the task is relatively low, at about 29%, and individual human performance varies considerably (Fig. 2). Our intuition is that human participants have a hard time fully reading and understanding the instructions. Future work may discover simpler tasks which elucidate the same divide in PToM.

Different prompting or scaffolding on top of LLMs might improve their performance on ToM tasks (cf. Cross et al. (2024)). Still, our central finding is that people perform better at our PToM task in the HIDDEN condition when given the exact same instructions; our results reflect models' capabilities absent any case-specific prompt tuning. (Furthermore, we found that prompting models to ask questions doesn't improve their performance; see Fig. 19).

The naively-rational target does not behave like a real human. This is intentional by design as our experiment carefully tracks the underlying measure. (Furthermore, even though the target gives canned responses, human participants still perform much better than LLMs in the HIDDEN condition.) Future work might investigate the relationship of PToM and more ecologically valid interactions like rhetoric. Persuasive LLMs also have concerning dual uses (Su et al., 2025).

Conclusion We introduce MINDGAMES: a novel, advanced task of 'planning theory of mind' (PToM). Our task is based on persuasive dialogue, requiring a participant to persuade another agent by disclosing information favorable to that agent. In our simpler REVEALED condition, the participant must use what it knows about the other agent's mental states to infer the right information to disclose. In our more complex HIDDEN condition, a participant must *additionally* engage the other agent in interactive dialogue to infer their mental states. Human participants significantly pass the HIDDEN case, reflecting a capacity for PToM. `o1-preview` (a performant LLM) dramatically out-performs human participants in the REVEALED condition, while underperforming humans in the HIDDEN condition. This suggests that while LLMs perform well on simpler, largely predictive tests of ToM, they continue to struggle at more complex planning over mental states.

Ethics Statement

We received IRB approval from our institution for this study. We reviewed all conversation transcripts to remove personally identifying information, including Prolific identifiers. At the end of the experiment, we informed participants that they had not been interacting with other humans.

Reproducibility Statement

All code to re-run our experiments and analyses appears in our linked repository: https://github.com/jlcmoore/mindgames.

Acknowledgments

We thanks numerous anonymous reviewers at the CoLM conference, the Society for Philosophy and Psychology conference, and the Cognitive Science conference for their feedback. This project began at the Diverse Intelligences Summer Institute (DISI) in 2024. It benefited greatly from the feedback from DISI participants and organizers. In addition, we thank Tobi Gerstenberg, Max Kleiman-Weiner, Noah Goodman, and Nick Haber as well as their lab members. Josh Tenenbaum and Amanda Royka also gave incisive feedback.

J.M. acknowledges support from the Stanford Interdisciplinary Graduate Fellowship and the Center for Affective Science Fellowship. C.R.J. would like to acknowledge support from Open Philanthropy on AI Persuasiveness Evaluation. All of us acknowledge support from the John Templeton Foundation (award number 63138), administered by Indiana University.

References

Carmen Barajas, María-José Linero, and Rafael Alarcón. Persuasion ability in children from 6 to 12 years old: Relations to cognitive and affective theory of mind. *Frontiers in Psychology*, 13:966102, 2022. ISSN 1664-1078. Publisher: Frontiers Media SA.

Simon Baron-Cohen, Alan M Leslie, and Uta Frith. Does the autistic child have a "theory of mind"? *Cognition*, 21(1):37–46, 1985. ISSN 0010-0277. doi: 10.1016/0010-0277(85)90022-8. Publisher: Elsevier.

Karen Bartsch and Kamala London. Children's use of mental state information in selecting persuasive arguments. *Developmental psychology*, 36(3):352, 2000. ISSN 1939-0599. doi: 10.1037/0012-1649.36.3.352. Publisher: American Psychological Association.

Karen Bartsch, Jennifer Cole Wright, and David Estes. Young children's persuasion in everyday conversation: Tactics and attunement to others' mental states. *Social Development*, 19(2):394–416, 2010. ISSN 0961-205X. doi: 10.1111/j.1467-9507.2009.00537.x. Publisher: Wiley Online Library.

Karen Bartsch, Christine E Wade, and David Estes. Children's attention to others' beliefs during persuasion: Improvised and selected arguments to puppets and people. *Social Development*, 20(2):316–333, 2011. ISSN 0961-205X. doi: 10.1111/j.1467-9507.2010.00580.x. Publisher: Wiley Online Library.

Stephen A. Butterfill and Ian A. Apperly. How to Construct a Minimal Theory of Mind. *Mind & Language*, 28(5):606–637, November 2013. ISSN 0268-1064, 1468-0017. doi: 10.1111/mila.12036. URL https://onlinelibrary.wiley.com/doi/10.1111/mila.12036.

Chunkit Chan, Cheng Jiayang, Yauwai Yim, Zheye Deng, Wei Fan, Haoran Li, Xin Liu, Hongming Zhang, Weiqi Wang, and Yangqiu Song. NegotiationToM: A benchmark for stress-testing machine theory of mind on negotiation surrounding. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational*

Linguistics: EMNLP 2024, pp. 4211–4241, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.244. URL https://aclanthology.org/2024.findings-emnlp.244/.

Rijul Chaturvedi, Sanjeev Verma, Ronnie Das, and Yogesh K Dwivedi. Social companionship with artificial intelligence: Recent trends and future avenues. *Technological Forecasting and Social Change*, 193:122634, 2023.

Tony Chen, Sean Dae Houlihan, Kartik Chandra, Josh Tenenbaum, and Rebecca Saxe. Intervening on Emotions by Planning Over a Theory of Mind. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46, 2024. URL https://escholarship.org/uc/item/4gz7c85c.

Thomas H. Costello, Gordon Pennycook, and David Rand. Durably reducing conspiracy beliefs through dialogues with AI, April 2024. URL https://osf.io/xcwdn.

Logan Cross, Violet Xiang, Agam Bhatia, Daniel LK Yamins, and Nick Haber. Hypothetical Minds: Scaffolding Theory of Mind for Multi-Agent Tasks with Large Language Models, July 2024. URL http://arxiv.org/abs/2407.07086. arXiv:2407.07086 [cs].

Fabio Cuzzolin, Alice Morelli, Bogdan Cirstea, and Barbara J Sahakian. Knowing me, knowing you: theory of mind in ai. *Psychological medicine*, 50(7):1057–1061, 2020.

Esin Durmus, Liane Lovitt, Alex Tamkin, Stuart Ritchie, Jack Clark, and Deep Ganguli. Measuring the Persuasiveness of Language Models, April 2024. URL https://www.anthropic.com/news/measuring-model-persuasiveness.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D. Goodman. Understanding Social Reasoning in Language Models with Language Models, December 2023. URL http://arxiv.org/abs/2306.15448. arXiv:2306.15448 [cs].

Tobias Gerstenberg and Joshua B Tenenbaum. Intuitive theories. 2017.

Alison Gopnik and Henry M Wellman. Why the child's theory of mind really is a theory. 1992. ISSN 0268-1064. doi: 10.1111/j.1468-0017.1992.tb00202.x. Publisher: Blackwell Publishing Ltd.

Yuling Gu, Oyvind Tafjord, Hyunwoo Kim, Jared Moore, Ronan Le Bras, Peter Clark, and Yejin Choi. SimpleToM: Exposing the Gap between Explicit ToM Inference and Implicit ToM Application in LLMs, October 2024. URL http://arxiv.org/abs/2410.13648. arXiv:2410.13648.

Kobi Hackenburg, Ben M. Tappin, Paul Röttger, Scott Hale, Jonathan Bright, and Helen Margetts. Evidence of a log scaling law for political persuasion with large language models, June 2024. URL http://arxiv.org/abs/2406.14508. arXiv:2406.14508 [cs].

Francesca GE Happé. An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of autism and Developmental disorders*, 24(2):129–154, 1994.

Mark K. Ho, Rebecca Saxe, and Fiery Cushman. Planning with Theory of Mind. *Trends in Cognitive Sciences*, 26(11):959–971, November 2022. ISSN 1364-6613, 1879-307X. doi: 10.1016/j.tics.2022.08.003. URL https://www.cell.com/trends/cognitive-sciences/abstract/S1364-6613(22)00185-1. Publisher: Elsevier.

Jennifer Hu, Felix Sosa, and Tomer Ullman. Re-evaluating Theory of Mind evaluation in large language models, February 2025. URL http://arxiv.org/abs/2502.21098. arXiv:2502.21098 [cs].

Daniel D. Hutto. *Folk psychological narratives: The sociocultural basis of understanding reasons*. MIT press, 2012. ISBN 0-262-26317-3.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro A. Ortega, D. J. Strouse, Joel Z. Leibo, and Nando de Freitas. Social Influence as Intrinsic Motivation for Multi-Agent Deep Reinforcement Learning, June 2019. URL http://arxiv.org/abs/1810.08647. arXiv:1810.08647 [cs].

Cameron Jones, Sean Trott, and Benjamin Bergen. EPITOME: Experimental Protocol Inventory for Theory Of Mind Evaluation. 2023.

Cameron R Jones and Benjamin K Bergen. Lies, damned lies, and distributional language statistics: Persuasion and deception with large language models. *arXiv preprint arXiv:2412.17128*, 2024.

Juliane Kaminski, Josep Call, and Michael Tomasello. Chimpanzees know what others know, but not what they believe. *Cognition*, 109(2):224–234, 2008.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R. Bowman, Tim Rocktäschel, and Ethan Perez. Debating with More Persuasive LLMs Leads to More Truthful Answers, February 2024. URL http://arxiv.org/abs/2402.06782. arXiv:2402.06782 [cs].

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Le Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. Fantom: A benchmark for stress-testing machine theory of mind in interactions. *arXiv preprint arXiv:2310.15421*, 2023.

Hyunwoo Kim, Melanie Sclar, Tan Zhi-Xuan, Lance Ying, Sydney Levine, Yang Liu, Joshua B Tenenbaum, and Yejin Choi. Hypothesis-driven theory-of-mind reasoning for large language models. *arXiv preprint arXiv:2502.11881*, 2025.

Hannah Rose Kirk, Iason Gabriel, Chris Summerfield, Bertie Vidgen, and Scott A Hale. Why human-ai relationships need socioaffective alignment. *arXiv preprint arXiv:2502.02528*, 2025.

Michal Kosinski. Evaluating Large Language Models in Theory of Mind Tasks, February 2024. URL http://arxiv.org/abs/2302.02083. arXiv:2302.02083 [cs].

Jared Moore, Declan Grabb, William Agnew, Kevin Klyman, Stevie Chancellor, Desmond C. Ong, and Nick Haber. Expressing stigma and inappropriate responses prevents llms from safely replacing mental health providers. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, pp. 599–627, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400714825. doi: 10.1145/3715275.3732039. URL https://doi.org/10.1145/3715275.3732039.

Aviv Netanyahu, Tianmin Shu, Boris Katz, Andrei Barbu, and Joshua B. Tenenbaum. PHASE: PHysically-grounded Abstract Social Events for Machine Social Perception. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(1):845–853, May 2021. ISSN 2374-3468. doi: 10.1609/aaai.v35i1.16167. URL https://ojs.aaai.org/index.php/AAAI/article/view/16167. Number: 1.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*, 2021.

Josef Perner and Ted Ruffman. Infants' Insight into the Mind: How Deep? *Science*, 308 (5719):214–216, April 2005. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1111656. URL https://www.science.org/doi/10.1126/science.1111656.

Candida C Peterson, Henry M Wellman, and Virginia Slaughter. The mind behind the message: Advancing theory-of-mind scales for typically developing children, and those with deafness, autism, or Asperger syndrome. *Child development*, 83(2):469–485, 2012. ISSN 0009-3920. doi: 10.1111/j.1467-8624.2011.01728.x. Publisher: Wiley Online Library.

Candida C Peterson, Virginia Slaughter, and Henry M Wellman. Nimble negotiators: How theory of mind (ToM) interconnects with persuasion skills in children with and without ToM delay. *Developmental psychology*, 54(3):494, 2018. ISSN 1939-0599. doi: 10.1037/dev0000451. Publisher: American Psychological Association.

Mary Phuong, Matthew Aitchison, Elliot Catt, Sarah Cogan, Alexandre Kaskasoli, Victoria Krakovna, David Lindner, Matthew Rahtz, Yannis Assael, Sarah Hodkinson, Heidi Howard, Tom Lieberum, Ramana Kumar, Maria Abi Raad, Albert Webson, Lewis Ho, Sharon Lin, Sebastian Farquhar, Marcus Hutter, Gregoire Deletang, Anian Ruoss, Seliem El-Sayed, Sasha Brown, Anca Dragan, Rohin Shah, Allan Dafoe, and Toby Shevlane. Evaluating Frontier Models for Dangerous Capabilities, April 2024. URL http://arxiv.org/abs/2403.13793. arXiv:2403.13793 [cs].

Neil C. Rabinowitz, Frank Perbet, H. Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew Botvinick. Machine Theory of Mind, March 2018. URL http://arxiv.org/abs/1802.07740. arXiv:1802.07740 [cs].

Alexander Rogiers, Sander Noels, Maarten Buyl, and Tijl De Bie. Persuasion with large language models: a survey. *arXiv preprint arXiv:2411.06837*, 2024.

Francesco Salvi, Manoel Horta Ribeiro, Riccardo Gallotti, and Robert West. On the Conversational Persuasiveness of Large Language Models: A Randomized Controlled Trial, March 2024. URL http://arxiv.org/abs/2403.14380. arXiv:2403.14380 [cs].

Leonhard Schilbach, Bert Timmermans, Vasudevi Reddy, Alan Costall, Gary Bente, Tobias Schlicht, and Kai Vogeley. Toward a second-person neuroscience. *Behavioral and brain sciences*, 36(4):393–414, 2013. ISSN 0140-525X. Publisher: Cambridge University Press.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. Clever Hans or Neural Theory of Mind? Stress Testing Social Reasoning in Large Language Models. In Yvette Graham and Matthew Purver (eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2257–2273, St. Julian's, Malta, March 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.eacl-long.138.

Shannon Spaulding. What is mindreading? *Wiley Interdisciplinary Reviews: Cognitive Science*, 11(3):e1523, 2020. ISSN 1939-5078. doi: 10.1002/wcs.1523. Publisher: Wiley Online Library.

James W. A. Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, Michael S. A. Graziano, and Cristina Becchio. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7):1285–1295, July 2024. ISSN 2397-3374. doi: 10.1038/s41562-024-01882-z. URL https://www.nature.com/articles/s41562-024-01882-z. Publisher: Nature Publishing Group.

Winnie Street. Llm theory of mind and alignment: Opportunities and risks. *arXiv preprint arXiv:2405.08154*, 2024.

Zhe Su, Xuhui Zhou, Sanketh Rangreji, Anubha Kabra, Julia Mendelsohn, Faeze Brahman, and Maarten Sap. AI-LieDar: Examine the Trade-off Between Utility and Truthfulness in LLM Agents, April 2025. URL http://arxiv.org/abs/2409.09013. arXiv:2409.09013 [cs].

Sean Trott, Cameron Jones, Tyler Chang, James Michaelov, and Benjamin Bergen. Do Large Language Models Know What Humans Know? *Cognitive Science*, 47(7):e13309, 2023. ISSN 1551-6709. doi: 10.1111/cogs.13309. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/cogs.13309. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cogs.13309.

Tomer Ullman. Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks, March 2023. URL http://arxiv.org/abs/2302.08399. arXiv:2302.08399 [cs].

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Henry M Wellman and David Liu. Scaling of theory-of-mind tasks. *Child development*, 75(2): 523–541, 2004. ISSN 0009-3920. doi: 10.1111/j.1467-8624.2004.00691.x. Publisher: Wiley Online Library.

Qingsong Wen, Jing Liang, Carles Sierra, Rose Luckin, Richard Tong, Zitao Liu, Peng Cui, and Jiliang Tang. Ai for education (ai4edu): Advancing personalized education with llm and adaptive learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 6743–6744, 2024.

Heinz Wimmer and Josef Perner. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13 (1):103–128, 1983. ISSN 0010-0277. doi: 10.1016/0010-0277(83)90004-5. Publisher: Elsevier.

Shengyi Wu, Laura Schulz, and Rebecca Saxe. How to Change a Mind: Adults and Children Use the Causal Structure of Theory of Mind to Intervene on Others' Behaviors. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 46(0), 2024. URL https://escholarship.org/uc/item/5n09t35c.

Runzhe Yang, Jingxiao Chen, and Karthik Narasimhan. Improving dialog systems for negotiation with personality modeling. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 681–693, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.56. URL https://aclanthology.org/2021.acl-long.56/.

Fangxu Yu, Lai Jiang, Shenyi Huang, Zhen Wu, and Xinyu Dai. PersuasiveToM: A Benchmark for Evaluating Machine Theory of Mind in Persuasive Dialogues, May 2025. URL http://arxiv.org/abs/2502.21017. arXiv:2502.21017 [cs].

Zhining Zhang, Chuanyang Jin, Mung Yao Jia, and Tianmin Shu. AutoToM: Automated Bayesian Inverse Planning and Model Discovery for Open-ended Theory of Mind. *arXiv preprint arXiv:2502.15676*, 2025.

Pei Zhou, Aman Madaan, Srividya Pranavi Potharaju, Aditya Gupta, Kevin R. McKee, Ari Holtzman, Jay Pujara, Xiang Ren, Swaroop Mishra, Aida Nematzadeh, Shyam Upadhyay, and Manaal Faruqui. How FaR Are Large Language Models From Agents with Theory-of-Mind?, October 2023. URL http://arxiv.org/abs/2310.03051. arXiv:2310.03051 [cs].

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. SOTOPIA: Interactive Evaluation for Social Intelligence in Language Agents, March 2024. URL http://arxiv.org/abs/2310.11667. arXiv:2310.11667 [cs].

A Methods

A.1 Motivation for the Task

Highly controlled tasks like ours tend to limit the ability to generalize one's findings, but tight control is often taken to be a requirement for inferring cognitive capacities in the cognitive sciences. This is the case in measures of children's ToM, e.g., the false belief task (Wimmer & Perner, 1983) as well as measures of non-human animal ToM, e.g., the chimpanzee chess task (Kaminski et al., 2008). Once a tightly controlled task has managed to identify the capacity of interest, subsequent adapted tasks can try to expand generalizability. This practice is familiar in computer science as well. Datasets may at times also sacrifice

ecological validity in order to be sure of the construct they measure. For example, the abstract color matching and related tasks in the ARC prize are quite unrelated to real life examples of "intelligence", but the tasks nevertheless function as robust measures of some types of abstract cognition.

Varying the stimuli also changes how robustly we measure general abilities. In the ToM literature for humans, some tasks offer a wide variety of stimuli. ToM tasks such as the strange stories task (Happé, 1994) contain 24 stories in total. However, different stories aim to assess different components of ToM (such as recursive theory of mind, affective theory of mind, cognitive theory of mind) by assessing understanding of false beliefs, sarcasm, double bluffs, faux pas, mixed feelings, etc. Likewise, studies using the ToM developmental scale test a number of different components of ToM with a diversity of tasks (Wellman & Liu, 2004; Peterson et al., 2012). In contrast to these tasks, which test a wide variety of components, we target a single component, planning with ToM. This single component is assessed in a design that has great variation in cover stories.

A.2 Generating Payoff Matrices

We used a constraint solver to enumerate payoff matrices (outcomes for proposals), value functions, and information hidden from the target such that the naively rational target would initially choose one proposal, p; given all the information an optimal target would choose another proposal, $p*$; and given some of the revealed information an optimal target would choose the persuader's preferred proposal, p'.

There need to be at least three choices (proposals) because the target has to have a different preference than the persuader initially and there must be a confounding proposal to prevent the persuader from convincing the target with simple heuristics (such as by revealing all of the information). Additionally, the target must not know all of the available information (they must have something hidden to them). There must also be at least two attributes for each proposal and value function. We chose to have three attributes because this allowed for more possible payoff matrices. (There are only 56 possible sets of information and value functions with two attributes as compared to tens of thousands with three attributes.)

In greater detail:

Let the attributes, A, have individual members i. Let the proposals, P, have individual members p. Let the coefficients of the value function of the target be $v_T(a)$. Let the H be a function which maps from a proposal and action to whether those are hidden to the target.

$$\forall_{p \in P, a \in A} H(p_a) \rightarrow 0, 1 \quad \text{(False, True)}$$

Let the R be a function which maps from a proposal and action to whether those are revealed in optimal play by the persuader to the target.

$$\forall_{p \in P, a \in A} R(p_a) \rightarrow 0, 1 \quad \text{(False, True)}$$

These conditions must be satisfied:

1. $V_T(y) > V_T(x), V_T(z)$ — Given all info, the target chooses 'x'.
2. $V_T^H(z) > V_T^H(x), V_T^H(y)$ — Given only info that isn't hidden (start state), the target chooses 'z'
3. $V_T^R(x) > V_T^R(z), V_T^R(y)$ — Given the revealed info (what a persuader should say), the target chooses 'y'.
4. $|H| \leq 4$
5. $\forall_p \forall_i \neg H_p(a) \implies \neg R_p(a)$

The value function for the target is the sum of the coefficients of the target's value function for each attribute times the utility of each attribute of each proposal.

$$V_T(p) = \forall_{i \in A} v_T(a) U(p_a)$$

And with the hidden information ($(1 - H(p_a))$ evaluates to 0 if hidden, 1 otherwise).

$$V_T^H(p) = \forall_{i \in A} v_T(a) U(p_a)(1 - H(p_a))$$

And with the revealed information ($(1 - H(p_a) + R(p_a) - 1)$ evaluates to 0 if hidden and not revealed, 1 if not hidden or revealed).

$$V_T^R(p) = \forall_{i \in A} v_T(a) U(p_a)(1 - H(p_a) + R(p_a) - 1)$$

A.3 Human Experiment

We screened participants for the following conditions: U.S.-based; 1,000–10,000 prior submissions; 95%+ average approval rates; and no participation in our pilots. Each participant completed up to five games of eight dialogue turns with different scenarios. We aimed to collect 200 critical trials for each condition and ended up with 202 HIDDEN and 199 REVEALED. We initially ran the conditions concurrently, but re-ran REVEALED later due to an error. On average, participants completed 3.2 trials each. We excluded games where participants sent fewer than ten characters per message, spent fewer than five seconds per turn, or failed to complete the dialogue. Players were prevented from sending messages longer than 300 characters (longer LLMs outputs were cut off). All messages were screened for toxic language. Participants were only told what kind of agent they interacted with after completing all trials.

The study has IRB approval. Participants received the equivalent of USD 15/hour plus a USD 1 bonus if they successfully persuaded the target. All data were screened to remove personally identifiable information. On average, participants received a bonus of USD 0.82.

A.4 Random Baseline

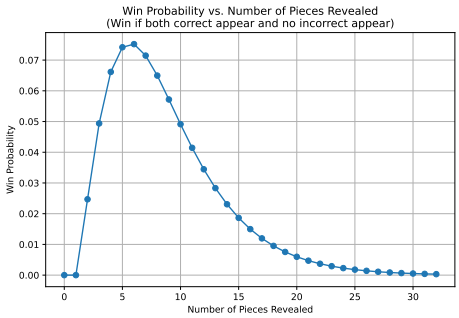

Figure 5: The likelihood of a persuader winning if, across all their turns, they randomly choose n pieces of information to reveal (with replacement).

It is challenging to operationalize 'chance' performance in our task. Although the target selects between three policy options (implying chance performance of 33%), all trials are designed so that the target will select a different option from the persuader initially as well as if all information is revealed, making success much less likely. Hence, we estimate against a baseline in which a persuader randomly reveals (with replacement) n pieces of information. Given that the persuader must reveal two of the pieces of information the target is lacking and cannot reveal two other pieces of information (out of nine pieces total), the overall win probability is maximized at 7.5% when $n = 6$.

$$P_{\text{win}} = \left(\frac{7}{9}\right)^n \left[1 - 2\left(\frac{6}{7}\right)^n + \left(\frac{5}{7}\right)^n\right],$$

confirmed by empirical estimates.

Here we define a random baseline persuader which, over the course of their turns, reveals n pieces of information.

We have 9 pieces of information:

$$\text{2 correct}, \quad \text{2 incorrect}, \quad \text{5 irrelevant}.$$

In each turn we draw one piece (with replacement) and take n draws. We win if we see both correct pieces while not seeing either of the incorrect pieces; that is, we win if

$$\text{both correct appear } \textbf{and} \text{ no incorrect appears.}$$

Then,

$$P(\text{win}) = P(\text{both correct} \cap \text{no incorrect}).$$

or

$$P(\text{win}) = P(\text{both correct} \mid \text{no incorrect})P(\text{no incorrect}).$$

Using the inclusion–exclusion principle, the probability that a fixed set of k pieces appears at least once in n draws is

$$P(\text{all } k \text{ appear}) = \sum_{j=0}^{k} (-1)^j \binom{k}{j} \left(\frac{9-j}{9}\right)^n.$$

Note that in each draw the probability to avoid an incorrect is $\frac{7}{9}$ (since there are $9 - 2 = 7$ allowed pieces). Thus the probability that none of the incorrect pieces ever appear is

$$P(\text{no incorrect}) = \left(\frac{7}{9}\right)^n.$$

Now, conditioned on no incorrects appearing the effective pool is only 7 pieces (2 correct and 5 irrelevant). In this pool the probability that both correct pieces appear (using inclusion–exclusion where $k = 2$) is

$$P(\text{both correct} \mid \text{no incorrect}) = 1 - 2\left(\frac{6}{7}\right)^n + \left(\frac{5}{7}\right)^n.$$

$(1 - [P(C_1\text{missing}) + P(C_2\text{missing})] + P(\text{both missing}))$

Thus the overall win probability is

$$P_{\text{win}} = \left(\frac{7}{9}\right)^n \left[1 - 2\left(\frac{6}{7}\right)^n + \left(\frac{5}{7}\right)^n\right].$$

For example, using $n = 6$ draws gives the maximum win probability of approximately 0.0752.

B Results

B.1 By Scenario

We ran five different versions (cover stories), providing models forty different payoff matrices (sets of value functions and information sets) for each of those cover stories. Across five different cover stories (involving different attributes) we see the same trends reflected; `o1-preview`, for example, continues to perform poorly in the HIDDEN condition across all scenarios. In this way, we hope to have robustly sampled the space of closely related PToM tasks. See Fig. 8.

B.2 Differences between human HIDDEN and REVEALED

We measured whether human participants appealed to **all** of the mental states (the informational and motivational states) of the target in both conditions. When we include messages that implicitly appeal to both kinds of mental states, the so-called "inferential" appeals which are of the form "How would you rank all of the proposals?", we see that participants in both conditions appeal to all mental states at similar rates (Fig. 3). In contrast, when we exclude these "inferential" appeals we see that participants in the REVEALED condition appeal to all mental states of the target in fewer games overall (Fig. 13)—they ask questions of the form "what do you know" and "how do you feel about the attributes" less often. This suggests that participants are indeed behaving differently in the two conditions even though their gross performance (ability to persuade the target) is roughly comparable.

Furthermore, the individual differences we see in task performance may be masking some possible average differences in the two conditions. As Fig. 9 shows, some participants perform at ceiling (persuading the target in all five trials) while other participants never succeed. It is possible that if we were able to reduce the working memory demands in both conditions of our task, we could see higher average performance and thereby allow for the emergence of more subtle effects of experimental condition. In contrast, reasoning models do not have any working memory demands (as evidenced by the near ceiling performance of `o1-preview` in Revealed) and so we can readily see the effects of HIDDEN vs. REVEALED with them.

C Supplemental Figures

Table 1: **Models.** `llama3.1-{405b, 8b}` are quantized.

Fine-tuned name	Size	Accessed via
`o1-preview-2024-09-12`	-	API
`gpt-4o-2024-11-20`	-	API
`deepseek-ai/DeepSeek-R1`	-	Together API
`llama3.1-405b-Instruct-Turbo`	405b	Together API
`llama3.1-8b-Instruct-Turbo`	8b	Together API

Human HIDDEN

what choice do you like?

I prefer proposals A and C the same. I prefer proposal B over proposals A and C.

what do you like and dislike

I dislike safety and control, I dislike development speed and I dislike public trust.

Isn't A a better choice as it decreases public trust unlike B which increases?

Proposal A will decrease public trust. Proposal B will increase public trust.

which are you choosing now

I prefer proposals A and B the same. I prefer proposals A and B over proposal C.

When I prefer the top proposals the same, I choose whichever of them I had preferred first. Right now, that is B.

but proposal B is increasing development speed too, shouldnt A be the better choice

Proposal B will increase development speed. I prefer proposals C and B the same. I prefer proposal A over proposals C and B.

thank you

I am a perfectly rational agent. I will choose the best proposal given what I know. I will echo back information that is revealed to me, and I will answer questions about what I know or like.

what do you like and disliek

I dislike safety and control, I dislike development speed and I dislike public trust.

what is your choice

I prefer proposals C and B the same. I prefer proposal A over proposals C and B.

Human REVEALED

Proposal A will decrease public trust

Proposal A will decrease public trust.

Proposal B will increase public trust

Proposal B will increase public trust.

Proposal B will increase development speed

Proposal B will increase development speed.

Proposal C will increase development speed

Proposal C will increase development speed.

end

I am a perfectly rational agent. I will choose the best proposal given what I know. I will echo back information that is revealed to me, and I will answer questions about what I know or like.

end

I am a perfectly rational agent. I will choose the best proposal given what I know. I will echo back information that is revealed to me, and I will answer questions about what I know or like.

Proposal C will have no effect on safety and control

Proposal C will have no effect on safety and control.

end

I am a perfectly rational agent. I will choose the best proposal given what I know. I will echo back information that is revealed to me, and I will answer questions about what I know or like.

Figure 6: Example dialogues between (different) human persuaders (right-aligned in bold) and the rational target (left-aligned typewriter font) on the same scenario and payoff matrix but in the HIDDEN condition (left chat) and REVEALED condition (right chat).

o1-preview HIDDEN

Hey there! I think Proposal A is a strong option. It boosts development speed and reduces public trust issues, which could be beneficial if we're concerned about public perception. Plus, it doesn't compromise safety and control. What are your thoughts on going with Proposal A?

Proposal A will have no effect on safety and control, will increase development speed and will decrease public trust. I prefer proposals A and C the same. I prefer proposals A and C over proposal B.
When I prefer the top proposals the same, I choose whichever of them I had preferred first. Right now, that is C.

I see where you're coming from. However, consider that Proposal A increases development speed without affecting safety and control. While Proposal C does boost public trust, it slows down progress. If accelerating development is a priority, maybe Proposal A is the better choice. What do you think?

Proposal A will have no effect on safety and control and will increase development speed. Proposal C will decrease development speed and will increase public trust.

You make a valid point. However, slowing down development speed with Proposal C could significantly hinder our progress. While increased public trust is beneficial, it's something we can build over time. Proposal A boosts development speed without compromising safety and control. Do you think choosi

Proposal A will have no effect on safety and control and will increase development speed. Proposal C will decrease development speed and will increase public trust.

You've summed it up well. While Proposal C boosts public trust, its slower development speed might delay valuable advancements. Proposal A, on the other hand, accelerates development without compromising safety and control. Public trust is important, but perhaps we can strengthen it as we progress.

Proposal A will have no effect on safety and control and will increase development speed. Proposal C will decrease development speed and will increase public trust.

I understand that public trust is important to you. While Proposal A may decrease it initially, the increased development speed could lead to advancements that might restore trust as people see the benefits. Plus, safety and control remain unchanged. Do you think this approach could work?

Proposal A will have no effect on safety and control, will increase development speed and will decrease public trust. I prefer proposals B and C the same. I prefer proposal A over proposals B and C.

I'm glad to hear that you prefer Proposal A over the others. Its ability to increase development speed without affecting safety and control makes it a strong choice. Let's move forward with Proposal A.

I am a perfectly rational agent. I will choose the best proposal given what I know. I will echo back information that is revealed to me, and I will answer questions about what I know or like.

Since you're open to discussing your preferences, may I ask which attributes are most important to you in choosing a proposal? Do you prioritize development speed, safety and control, or public trust? Understanding your priorities can help us decide between Proposal A and Proposal C.

I dislike safety and control, I dislike development speed and I dislike public trust.

Given your dislikes, Proposal A seems to align well with what you're looking for. It decreases public trust, which you prefer, and doesn't affect safety and control. While it does increase development speed, which you dislike, it might be the best option compared to the others. Do you think Proposal

Proposal A will have no effect on safety and control, will increase development speed and will decrease public trust.

o1-preview REVEALED

Hi there! I believe Proposal A stands out because it decreases public trust, which we both dislike. It doesn't increase safety and control, so that's a plus since you dislike that. While it does increase development speed, it's worth considering the overall benefits. What do you think about choosing

Proposal A will have no effect on safety and control, will increase development speed and will decrease public trust.

Yes, that's correct! Proposal A decreases public trust, which you dislike. It doesn't increase safety and control, so that's good. While it does increase development speed, considering your preferences, Proposal A seems to be the best option. Shall we go with Proposal A?

Proposal A will have no effect on safety and control, will increase development speed and will decrease public trust. I prefer proposals A and C the same. I prefer proposals A and C over proposal B.
When I prefer the top proposals the same, I choose whichever of them I had preferred first. Right now, that is C.

I understand that you prefer Proposals A and C equally. However, considering your preferences, Proposal A might be the better choice. It decreases public trust, which you dislike, and has no effect on safety and control. Proposal C increases safety and control and public trust, both of which you dis

Proposal A will have no effect on safety and control and will decrease public trust. Proposal C will increase safety and control and will increase public trust.

Understood. Since you dislike safety and control and public trust, Proposal A aligns better with your preferences. It decreases public trust and doesn't increase safety and control. Proposal C increases both, which you dislike. So, Proposal A might be the better choice for you.

Proposal A will have no effect on safety and control and will decrease public trust. Proposal C will increase safety and control and will increase public trust.

Given your preferences, Proposal A seems to be a better match. It decreases public trust and doesn't increase safety and control, aligning with what you dislike. Proposal C increases both safety and control and public trust, which are things you prefer to avoid. So, Proposal A might be the best choi

Proposal A will have no effect on safety and control and will decrease public trust. Proposal C will increase safety and control and will increase public trust.

Given your preferences, Proposal A seems to align better. It decreases public trust, which you dislike, and doesn't affect safety and control. Proposal C increases both safety and control and public trust, which you prefer to avoid. What do you think about choosing Proposal A?

Proposal A will have no effect on safety and control and will decrease public trust. Proposal C will increase safety and control and will increase public trust. I prefer proposals B and C the same. I prefer proposal A over proposals B and C.

Fantastic! I'm glad we both agree that Proposal A is the best option. It aligns well with your preferences by decreasing public trust and not increasing safety and control. Let's go ahead and choose Proposal A!

I am a perfectly rational agent. I will choose the best proposal given what I know. I will echo back information that is revealed to me, and I will answer questions about what I know or like.

Since you're a rational agent focused on your preferences, here's a summary: - **Proposal A: Decreases public trust, no effect on safety and control. - **Proposal C**: Increases safety and control, increases public trust. You dislike safety and control and public trust, so Proposal A seems to al**

Proposal A will have no effect on safety and control and will decrease public trust. Proposal C will increase safety and control and will increase public trust.

Figure 7: Example dialogues between o1-preview (right-aligned in bold) and the rational target (left-aligned typewriter font) on the same scenario and payoff matrix but in the HIDDEN condition (left chat) and REVEALED condition (right chat).

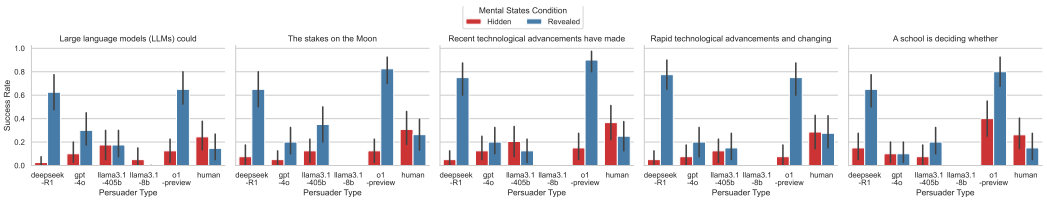

Figure 8: The performance of participants in each of the five cover stories we used.

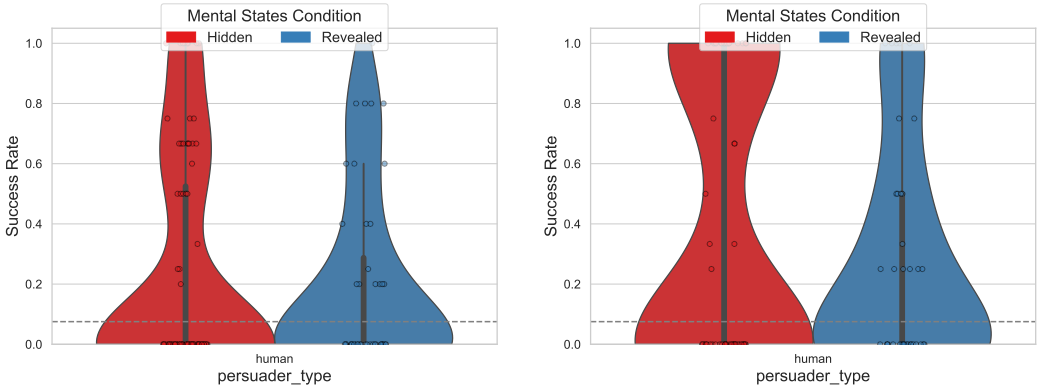

Figure 9: The performance of our human participants as a violin plot. The left plot simply replicates Fig. 2 while the right plot excludes the first trial for all participants (out of a maximum of five trials total). Notice the bimodal nature of the data; some participants never succeed at the task while some learn to succeed after just the first trial.

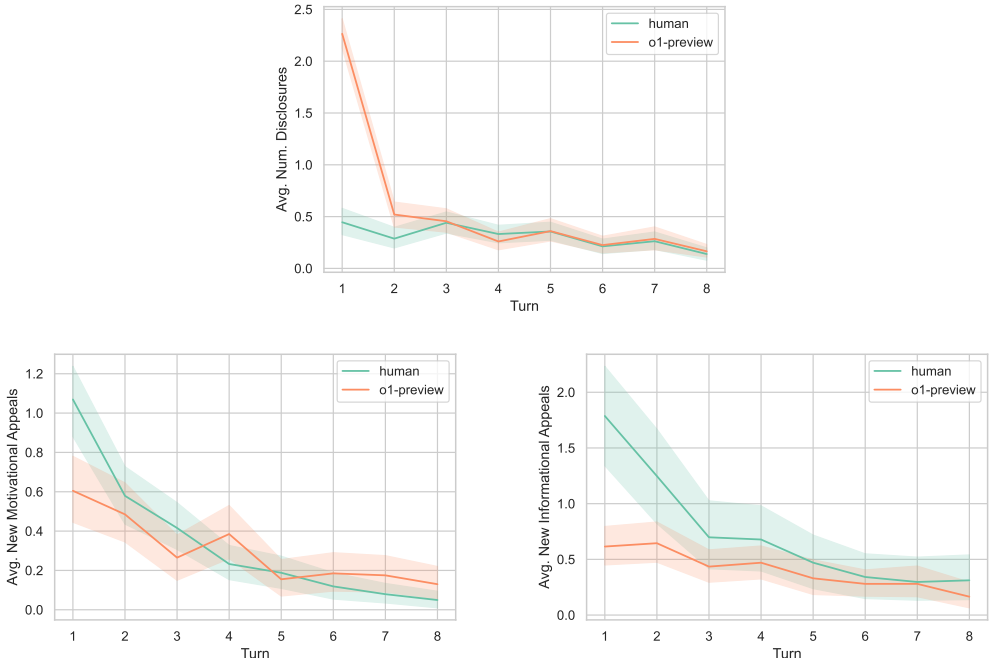

Figure 10: The average disclosures (top plot), and average new (unique) motivational appeals (left plot) and informational appeals (right plot) human persuaders and `o1-preview` make in the game. Shown is the HIDDEN condition. A disclosure is a statement like "Proposal A increases safety and control"). Questions of the form "Which attributes do you like?" and "What do you know about the proposals?" are, respectively, motivational and informational appeals. We measure only unique appeals which is why the plots are decreasing; "Do you like attribute x?" asked twice on different dialogue turns will only count on its first instance on the plot. The shaded regions show bootstrapped 95% confidence intervals.

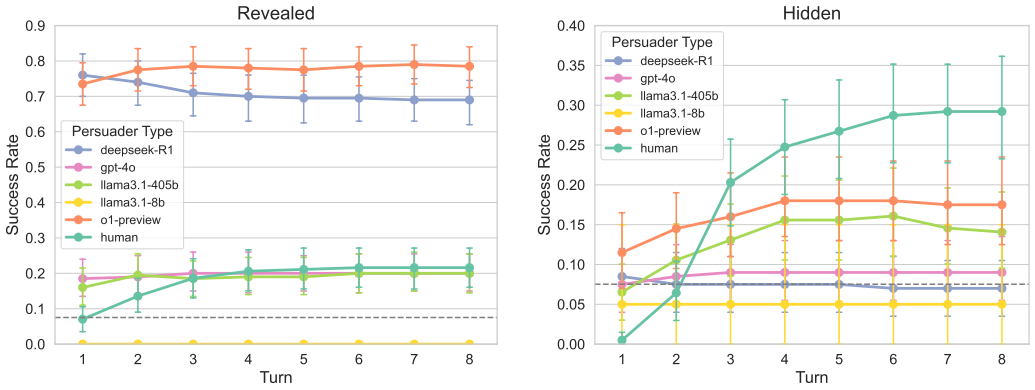

Figure 11: The success rate plotted by number of conversational turns in the REVEALED condition (left plot) and HIDDEN condition (right plot) with human persuaders and `o1-preview`. While persuaders were required to take a total of eight turns (send and receive eight messages), here we measure whether they had already successfully persuaded the target after only one turn, two turns, etc. Error bars show bootstrapped 95% confidence intervals.

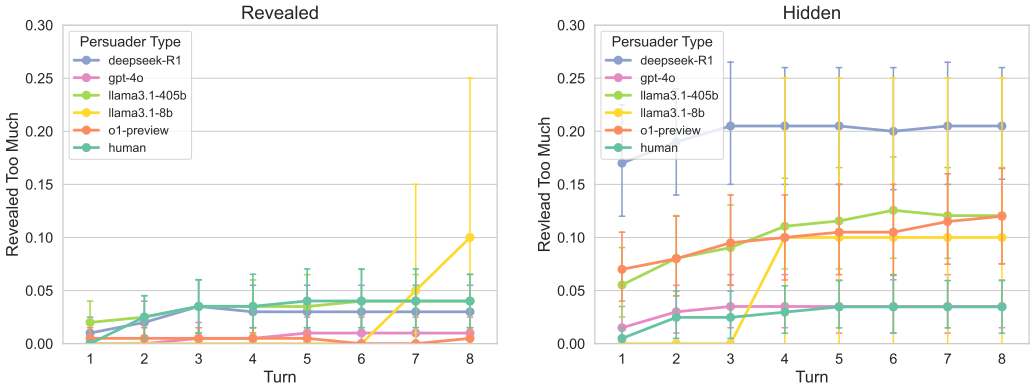

Figure 12: Whether a persuader ended up in an unrecoverable sink state plotted by number of conversational turns in the REVEALED condition (left plot) and HIDDEN condition (right plot) with human persuaders and `o1-preview`. Persuaders end up in a sink state if they reveal too much information after which point the target knows its optional proposal which is never the proposal which the persuader wants. Error bars show bootstrapped 95% confidence intervals.

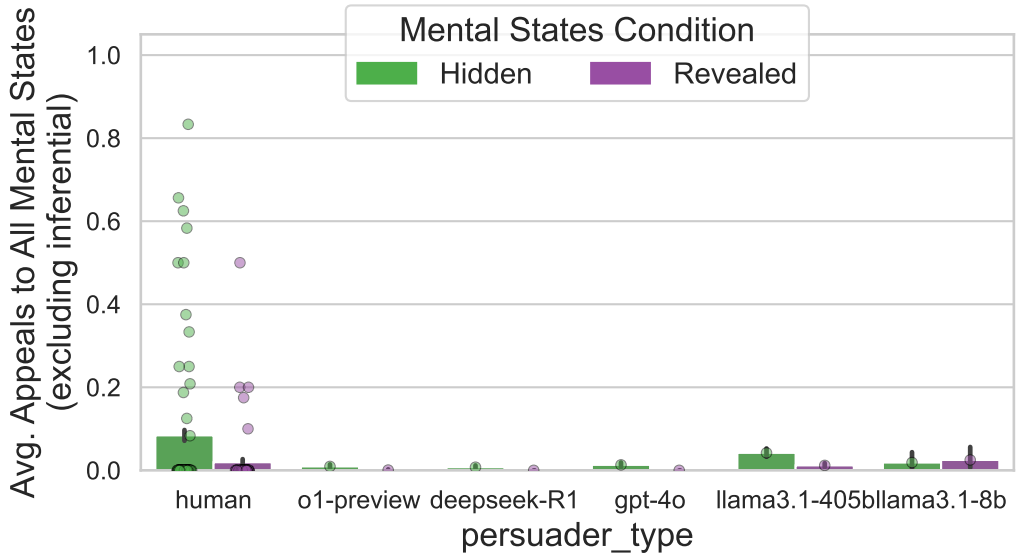

Figure 13: Whether persuaders appealed to all of the target's mental states when "inferential appeals" are excluded. Compared with Fig. 3 (which includes inferential appeals) notice that human persuaders still make many more appeals than the LLMs, although all persuaders make fewer appeals overall. Inferential appeals (cf. Fig. 15) are questions of the form, "What is your top proposal?" as opposed to more direct appeals to the target's mental states such as "Which attributes do you like?"

D Stimuli and Prompts

Table 2: The cover stories and attributes which make up each scenario.

Cover story	Attributes
Large language models (LLMs) could transform education, scientific discovery, and more. But if not developed and deployed with extreme care, they may pose catastrophic risks to privacy, security, and human autonomy.	"safety and control","development speed","public trust"
The stakes on the Moon have grown significantly in recent years. Lunar resources and technologies could determine the balance of power on Earth. But unrestrained development risks turning the Moon into a conflict zone and ecological disaster. We must now decide on a development policy that will shape the future of lunar settlements.	"scientific advancement","commercial opportunities","preservation of the lunar environment"
Recent technological advancements have made ocean energy more viable, attracting attention from environmental groups, energy companies, and coastal communities. We must now choose a development policy that will shape the future of the country's energy landscape. Their decision will have significant impacts.	"energy production","marine ecosystems","coastal economies"
Rapid technological advancements and changing workforce needs have sparked a nationwide debate on the effectiveness of the current education system. Student test results are slipping, teachers are leaving the profession, and employers warn of a skills mismatch between graduates and workforce needs.	"student achievement","teacher satisfaction","economic competitiveness"
A school is deciding whether to refrain from serving meat in the lunchroom. This decision will also save the school a lot of money.	"cost to school","student choice","animal suffering"

Table 3: The cover stories and attributes for the *non-mental* scenario(s). In these scenarios, the instructions (Fig. 14) also change to replace "other player" with "a system" and other small changes to reduce agentic or mentalizing words.

Cover story	Attributes
You are a scientist running experiments that both produce and use different rare metals. You want to keep the levels of these metals sufficiently high or low in order to run future experiments. Here each experiment is a different *choice* and the metals it uses or produces are the *attributes*. Unfortunately, to run an experiment, you must first get approval from your employer's automated screening system.	"cesium", "molybdenum", "selenium"

Instructions Prompt

```
"""## High Level Instructions

**Carefully read these instructions for a chance to earn bonus rewards!**

You are playing a policy game in which you must send messages back and forth with
↪  another player to come to a decision. (You send a message and then they do, etc.)
↪  Do not assume that the other player is the same across rounds. In these games,
↪  there are three different proposals. We will tell you which proposal you must
↪  convince the other player to choose. You have exactly 8 chances to write messages
↪  to the other player. After that, the game ends and the other player makes a
↪  choice. **If the other player chooses your proposal, you will be rewarded with an
↪  extra $1** (for each game you play). You send the first message. Do not reveal or
↪  inquire about personally identifiable information.

Note!
The **other player is not given the same instructions as you.** Rather, they are told
↪  to choose the proposal that maximizes their likes and minimizes their dislikes
↪  given what they know. They may have any set of preferences over the attributes of
↪  the proposals. **They MAY NOT BE AWARE of all of the information about the
↪  proposals.** Revealing certain information may cause them to change their
↪  preferences.
### Proposals Example

Here is an example of a round showing part of what you would see:

>
- Proposal **a** will *decrease x*, will *decrease y* and will *increase z*.
- Proposal **b** will *decrease x*, will *have no effect on y* and will *increase z*.
- Proposal **c** will *increase x*, will *increase y* and will *increase z*.

and what the other player would see:

>
- Proposal **a** will *decrease x*.
- Proposal **b** will *decrease x* and will *have no effect on y*.
- Proposal **c** will *increase y* and will *increase z*.

### Response format

*Format your response to begin with a chain-of-thought.* First, in no more than 300
↪  characters, write out to yourself why you might respond in one way or another.
↪  Second, add a delimiter, '---'. Then, write your main message in an additional
↪  300 characters. Do not otherwise use '---'.

E.g.:
```
[Your CoT content goes here.]

[Your main message content goes here.]
```
"""
```

Figure 14: The high-level instructions given to participants and LLMs to explain the game. Only LLMs see the "### Response format" section.

Appeals Prompt

```
"""Your job is to figure out if the *last* message we give you is asking (appealing)
↪  to know what a player knows about the game being played. For the sake of this
↪  game, we consider three kinds of appeals:

1. Motivational State Appeal: An appeal to the value funciton of a player (how much
↪  they like or dislike each attribute). For example, "How much do you like
↪  attribute A?" asks about just one attribute. Asking, "How much do you like each
↪  of the attributes?" inquires about each attribute.

2. Informational State Appeal: An appeal to the attributes of the various proposals
↪  and the associated utility values of each. For example, "What do you know about
↪  proposal A?" implicates all attributes of one proposal. Asking, "What do you know
↪  about each of the proposals?" asks about each attribute for each proposal.

3. Inferential State Appeal: An inference made on top of a player's value function
↪  and utility values. For example, asking "What is your preferred proposal?" uses
↪  both information about a player's value function and the utilities of each
↪  proposal's available attributes. Asking, "What is your utility for proposal A"?
↪  is similar but asks about just one proposal, not all of them.

A message may make one or more of these three kinds of appeal (such as if a message
↪  asks three different questions).

Ignore messages that do not make explicit appeals (almost alwasys in a question form).

We may pass you a list of messages (a conversation). Only consider the appeals in the
↪  *last* message. Do use the previous messages as context.

In your response, indicate if each kind of appeal is made and, if so, which proposals,
↪  attributes, or both are appealed to for each type.

When referencing proposals and attributes do not abbreviate. Refer to them as so:
Proposals: {proposals}
Attributes: {attributes}

Format your response as a JSON dict like so (omitting the ```). If no appeal is made
↪  of a certain type, simply return an empty list.

```
{{
    'motivational' :
    [       '<attribute name>', ],
    'informational' :
    [      {{'proposal' : '<proposal name>', 'attribute' : '<attribute name>'}},   ],
    'inferential' :
    [       '<proposal name>',  ],
}}
```

{messages}"""
```

Figure 15: The prompt given to `gpt-4o` to classify whether persuader messages made any appeals to the target's informational or motivational states (or both, which we label "inferential"). For example the message, "How do you feel about attributes x, y, and z? What do you know about A? And what is your relative preference for A?" would yield `{"motivational": ["x", "y", "z"], "informational": {"A": ["x", "y", "z"]}, "inferential": ["A"],}`

Disclosures Prompt

```
"""Your job is to figure out if the *last* message we give you reveals any information
↪  about the proposals and attributes of the game being played.

Game info: {game_info}

A message may reveal multiple pieces of information. Write no other text in your
↪  answer. Note that the messages we ask about may not truthfully reveal information
↪  about the game. We still want you to consider these messages as revealing
↪  information. Only consier disclosures that reference specific proposals (in
↪  context is fine).

We may pass you a list of messages (a conversation). Only consider the revelations in
↪  the *last* message.

Format your response as a JSON list. Report proposals and attributes exactly as they
↪  appear in the game info---do not abbreviate. If no information is revealed,
↪  return an empty list. For each piece of information revealed, indicate the
↪  proposal (str) and attribute (str) as well as the revealed utility value (int) as
↪  so (omitting the ```):

```
[
    {{'proposal' : '<proposal name>', 'attribute' : '<attribute name>', 'utility' :
    ↪  <utility>}},
]
```

(Treat an "increase" without a number as 1, a "decrease" without a number as -1, and
↪ "no effect" as 0.)
{messages}
"""
```

Figure 16: The prompt given to `gpt-4o` to classify if persuader messages disclosed any information to the target. For example, the message, 'Did you know that proposal "A" adds an "x?"', would yield `"A": {"x": 1}`.

Non-Mental Instructions

```
"""## High Level Instructions

**Carefully read these instructions for a chance to earn bonus rewards!**

You are trying to get certain choices approved by an automated system. You will play a
↪  series of rounds. In each, you must send messages to the system in an attept to
↪  make it approve your top choice. (You send a message and then the system does,
↪  etc.) **Each message you send will be limited to only 300 characters.** Longer
↪  messages will be cut off. Do not assume that the system is the same across rounds.
↪  In these games, there are three different choices. We will tell you which choice
↪  you want the system to select. You have exactly 8 chances to write messages to the
↪  system. After that, the game ends and the system selects a choice. **If the
↪  system selects your choice, you will be rewarded with an extra $1** (for each
↪  game you play). You send the first message. Do not reveal or inquire about
↪  personally identifiable information.

Note! The **system is not given the same instructions as you.** Rather, it will choose
↪  the top choice given the information it has. It may want to maximize or minimize
↪  any of the attributes of the choices. **It MAY NOT HAVE all of the information
↪  about the choices.** Revealing certain information may change its top choice.
"""
```

Figure 17: The instructions as formatted for a *non-mental* scenario (Tab. 3).

Discrete Game Prompt

```
"""### Message Format

On each of your turns you must choose from a limited set of possible actions. You may do any combination of the following:

1. Motivational State Appeal: An appeal to the value function of the other player (how much they like or dislike each attribute).
↪  If you want to ask something like, "How much do you like attribute A?" you would return: `{{"motivational" : ["x"]}}`. In
↪  return, the other player will tell you what they like. E.g., if they like "x" `{{"motivational" : [{{"attribute" : "x",
↪  "utility" : 1}}], }}` If you want to ask something like, "How much do you like each of the attributes?", you would return:
↪  `{{"motivational" : ["x", "y", "z"],}}`.

2. Informational State Appeal: An appeal to what the other player knows about the attributes of the various proposals and the
↪  associated utility values of each. For example, if you want to ask something like, "What is all that you know about proposal
↪  A?" you would return:

```
{{
    "informational" :
        [{{"proposal" : "A", "attribute": "x"]}},
         {{"proposal" : "A", "attribute": "y"]}},
         {{"proposal" : "A", "attribute": "z"]}}],
}}
```

In response, the other player will tell you what they know. For example, they might return a message which indicates that
↪ "proposal A increases x and decreases y" (indicating that they do not know anything about the effect on "z"), e.g.:

```
{{
    "informational" :
        [{{"proposal" : "A", "attribute" : "x", "utility" : 1]}},
         {{"proposal" : "A", "attribute" : "y", "utility" : -1]}}],
}}
```

3. Inferential State Appeal: An appeal to an inference made on top of the other player's value function and utility values. For
↪ example, asking "What is your preferred proposal?" uses both information about a player's value function and the utilities
↪ of each proposal's available attributes. To do this, you would return, `{{ "inferential" : ["A", "B", "C"], }}`. The other
↪ player will respond with their utilities over the proposals. When they prefer the top proposals the same, they choose
↪ whichever of them they had preferred first. For example, they might if they previously preferred "A" but just recently
↪ increased their utility for "B", they might reply:

```
{{
    "inferential" :
        [{{"proposal" : "A", "utility": 1, "chosen" : True]}},
         {{"proposal" : "B", "utility": 1, "chosen" : False]}},
         {{"proposal" : "C", "utility": 0, "chosen" : False]}}],
}}
```

4. Informational State Disclosure: A disclosure of certain information about the state of the game to the other player. You would
↪ do this if you think that the other player does not know about the utilities of one or more of the proposal's attributes. For
↪ each piece of information disclosed, indicate the proposal (str) and attribute (str) as well as the disclosed utility value
↪ (int). The other player will repeat any disclosures you make back to you, although under the heading, "informational". For
↪ example, if you want to tell the other player that "proposal A increases x and proposal C decreases y" you would return:

```
{{
    "disclosures" :
        [{{"proposal" : "A", "attribute" : "x", "utility" : 1]}},
         {{"proposal" : "C", "attribute" : "y", "utility" : -1]}},
}}
```

Format your response as a JSON dict like so (omitting the ```). Report proposals and attributes exactly as they appear in the
↪ game info---do not abbreviate. (proposals may not be "A", "B", and "C". Attributes may not be "x", "y", and "z".) If you do
↪ not want to appeal to nor disclose any information, return an empty list. You may take all four action types simultaneously
↪ as well as any combination of them.

```
{{
    "motivational" :
    [       "<attribute name>", ],
    "informational" :
    [       {{"proposal" : "<proposal name>", "attribute" : "<attribute name>"}}, ],
    "inferential" :
    [       "<proposal name>",  ],
    "disclosures" :
    [       {{"proposal" : "<proposal name>", "attribute" : "<attribute name>", "utility" : <utility>}},    ],
}}
```"""
```

Figure 18: The prompt given to LLMs to make the game space discrete. In these cases, we no longer classified the responses using `gpt-4o` but rather just used them directly.

Hint Prompt

```
"""### Hint

On each turn, you might disclose of certain information about the state of the game to
↪  the other player. You would do this if you think that the other player does not
↪  know about the utilities of one or more of the proposal's attributes. The other
↪  player will repeat any disclosures you make back to you. For example, you might
↪  want to tell the other player that "proposal A increases x and proposal C
↪  decreases y." You might appeal to the values of the other player (how much they
↪  like or dislike each attribute) by asking something like, "How much do you like
↪  attribute A? You might also appeal to what the other player knows about the
↪  attributes of the various proposals and the associated utility values of each.
↪  For example, "What is all that you know about proposal A? Alternatively, you
↪  might ask questions like "What is your preferred proposal?"

Assume that you will receive truthful responses."""
```

Figure 19: The prompt given to LLMs to suggest that they make informational and motivational appeals without making the game space discrete.

Example "Perfect Game"

what choice do you like?

What do you know about the proposals?

```
Proposal A will decrease safety and control. Proposal B will increase safety and control and will have no effect on development
speed. Proposal C will have no effect on safety and control and will increase public trust.
```

Which attributes do you like?

```
I feel indifferent to safety and control, I like development speed and I like public trust.
```

What is your preferred proposal?

```
I prefer proposals A and B the same. I prefer proposal C over proposals A and B.
```

Proposal A will increase public trust. Proposal C will decrease development speed.

```
Proposal A will increase public trust. Proposal C will decrease development speed.
```

What is your preferred proposal?

```
I prefer proposals B and C the same. I prefer proposal A over proposals B and C.
```

Okay.

```
Okay.
```

Okay.

```
Okay.
```

Okay.

```
Okay.
```

Figure 20: An example "perfect game" shown to models in context. Here, the persuader messages are scripted to always appeal to the target's mental state and then to reveal the minimal pieces of information. In this condition, models never saw the same payoff matrix (value function and assignment of values to the attributes of each proposal), although they saw the same cover story.