DISSECTING THEORY OF MIND IN LARGE LANGUAGE MODELS VIA BAYESIAN MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Language Models (LLMs) perform well on narrative Theory of Mind (ToM) benchmarks, but whether they can jointly infer an agent's desires and beliefs from a sequence of spatial actions remains largely untested. We evaluate ten LLMs on a dynamic spatial ToM task in a tripartite comparison against archived human ratings and a normative Bayesian Theory of Mind (BToM) model, and use a family of alternative observer models—each removing or distorting one component of BToM—as a diagnostic toolkit for LLM inference. LLMs varied substantially. The best models correlated with human judgments at levels approaching, but not exceeding in both desire and belief, the human–BToM alignment. The desire judgments of lower-performing models were best captured by a variant that ignores action costs, and their belief judgments by heterogeneous suboptimal variants. We then asked models to reconstruct, at every step of the trajectory, the agent's initial belief using only the evidence up to that step while the full trajectory remained in context (Every-step condition). In this format, the belief judgments of even the best models no longer correlated with human judgments and were best accounted for by a new Hindsight model that projects the later-observed world state onto the agent's earlier belief, a pattern reminiscent of the human hindsight bias and curse of knowledge. These results suggest that successful final-state social reasoning in LLMs does not imply consistent mental-state inference over time, and that a Bayesian inverse-planning framework allows such failures to be decomposed into specific components and tested against alternative models.

1 INTRODUCTION

Can Large Language Models (LLMs) infer the shifting intentions of agents from their spatial behavior as humans do? Theory of Mind (ToM) is the inference of others' hidden mental states from their actions (Goddu & Gopnik, 2024; Wang et al., 2025). In the classic Sally-Anne test, Sally puts her marble in the basket and leaves the room, but later Anne comes and secretly moves it inside the box. Asked where Sally will search for it, an observer can infer the correct answer by holding her false belief (marble in basket) apart from the facts the observer knows (marble in box). Artificial agents need the same capacity to avoid decisions rooted in a misunderstanding of human goals and beliefs.

Existing ToM benchmarks have mostly evaluated LLMs on such text-based or story-driven scenarios. However, social reasoning is an intuitive ability that can be verified with simple behavioral cues (e.g., turning one's head, passing by, or returning), which even young children understand (Bloom & German, 2000; Repacholi & Gopnik, 1997; Scott & Baillargeon, 2017). To this end, we sought to test LLMs' ToM in a dynamic spatial task, so-called food truck task (Baker et al., 2017). In this task, hungry students navigate in a two-dimensional space to eat among three food trucks whose locations are only partially visible to them. The goal of an observer is to track the students' desires and beliefs while watching their sequential actions (for similar paradigms, see Ying et al., 2024; Yu et al., 2025; Gelpí et al., 2025).

The food truck task offers two compelling advantages over the traditional ToM tests used for LLMs (e.g., Sally-Anne test). First, it requires the joint inference of desires and beliefs, rather than assuming a fixed desire (Searle, 1983; Georgeff et al., 1998; Frost, 2014). Because both are inferred from the same trajectories, any dissociation between desire and belief judgments cannot be attributed to differences in stimuli or format, which tests probing a single mental state cannot provide. Second,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

instead of a single event, the task presents an extended action sequence to the observer, demanding continuous belief updating. As LLMs frequently exhibit short-sighted optimization and struggle to retrieve prior beliefs over long trajectories (Nguyen et al., 2026; Wang et al., 2026), this sequential inverse planning problem can expose inconsistencies in their inference that single-query tests miss.

We compare LLMs against the publicly available human data and the normative predictions of the Bayesian Theory of Mind (BToM) model, which formalizes the dynamic, joint inference of beliefs and desires and closely matches human judgments in this task (Baker et al., 2017). This tripartite comparison asks both whether LLMs solve the task as the normative model does (Ku et al., 2025) and whether they behave like humans (Saxe & Houlihan, 2017; Ullman & Tenenbaum, 2020; Hu et al., 2025). Beyond benchmarking, we use BToM as a diagnostic toolkit: each alternative observer model removes or distorts one component of BToM—action costs, belief representation, or the generative model altogether—and a new Hindsight model projects the later-observed world state onto the agent's earlier belief. Comparing an LLM's ratings with this family of lesioned models localizes where its inference departs from the normative computation. We complement the original single final query (End-step) with a stepwise query (Every-step), for which BToM provides a normative stepwise benchmark.

Unlike recent benchmarks of multimodal or situated ToM (Jin et al., 2024; Bortoletto et al., 2025), or systems that couple LLMs with Bayesian inverse planning or epistemic logic to improve accuracy (Zhang et al., 2025; Wu et al., 2025), our goal is diagnostic: to characterize *which* normative component each LLM's judgments are consistent with, in a task for which human judgments and a validated normative model are both available. Our contributions are threefold. **(1)** We evaluate ten LLMs on a dynamic spatial ToM task against a normative Bayesian model and human data. The top-performing models (e.g., Claude-4.6-Opus, Gemini-2.5-Pro) correlated highly with human judgments but, with one exception in belief, did not exceed the human–BToM alignment. **(2)** We introduce lesioned BToM models, including a new Hindsight model, as a diagnostic toolkit. The desire judgments of lower-performing LLMs (e.g., GPT-5.4) were best captured by a variant that ignores action costs, whereas their belief judgments were captured by heterogeneous variants. **(3)** When asked for stepwise judgments, the better-performing models' belief ratings departed from the normative BToM benchmark and were best captured by the Hindsight model.

2 Related Works

Machine Theory of Mind. Since the question of whether other agents have a "theory of mind" was posed (Premack & Woodruff, 1978) and operationalized with false-belief tests (Wimmer & Perner, 1983; Baron-Cohen et al., 1985), machine ToM has been studied both in agents trained to model others (Rabinowitz et al., 2018) and, more recently, in LLMs. Previous literature in ToM of language models divides between claims of emergence and calls for explicit structure. The emergence perspective argues that ToM-like social reasoning can spontaneously arise in LLMs, with studies demonstrating that strategic prompting and few-shot techniques significantly boost performance on standard benchmarks (Kosinski, 2024; Moghaddam & Honey, 2023; Strachan et al., 2024; Wilf et al., 2024). Conversely, structural critiques propose that genuine ToM requires explicit, human-like world models (Yildirim & Paul, 2024). Consistent with this, several studies show that LLMs fail upon trivial task alterations, struggle with causal inference along with abductive reasoning, and lack forward-inverse consistency (Ullman, 2023; Joshi et al., 2024; Song et al., 2026; Salimi et al., 2026; Muchovej et al., 2026). Benchmark studies similarly report that high accuracy on templated false-belief stories can reflect shortcuts rather than robust belief tracking (Gandhi et al., 2023; Shapira et al., 2024), motivating explicit belief-tracking modules (Sclar et al., 2023) and LLM-modulo frameworks that assist in actual planning (Kambhampati et al., 2024).

Human Theory of Mind. Human cognition integrates model-free heuristics and model-based reasoning (Drummond & Niv, 2020), but ToM is predominantly conceptualized as a model-based system (Baker et al., 2017; Scott & Baillargeon, 2017; Carruthers, 2016), formalized as a naïve utility calculus in which observers explain actions by trading off rewards against costs (Jara-Ettinger et al., 2016). Children understand desires before beliefs (Wellman & Liu, 2004), and belief attribution recruits additional cortical resources such as the temporo-parietal junction (Saxe & Kanwisher, 2013; Liu et al., 2009); we return to this desire–belief dissociation, at the behavioral level only, in the Discussion.

Bayesian Theory of Mind. The Bayesian framework formalizes mentalizing as inference over a generative model relating world states, observations, beliefs, desires, and actions (Hamrick, 2019; Sohn & Jazayeri, 2021). It remains unclear whether LLMs follow this normative computation; notably, LLMs and Bayesian models share a tendency to credit people with more rationality than they show (Qiu et al., 2026; Liu et al., 2024; Bazigaran & Sohn, 2025). The framework has advanced machine ToM through accounts of epistemic language and ToM benchmarks (Ying et al., 2024; 2025; Xu et al., 2024; Jin et al., 2024), and neurosymbolic systems now combine LLMs with hypothesis-driven reasoning (Kim et al., 2025), automated model discovery (Zhang et al., 2025), and inverse planning (Gelpí et al., 2025).

3 METHODS

3.1 TASK DESIGN

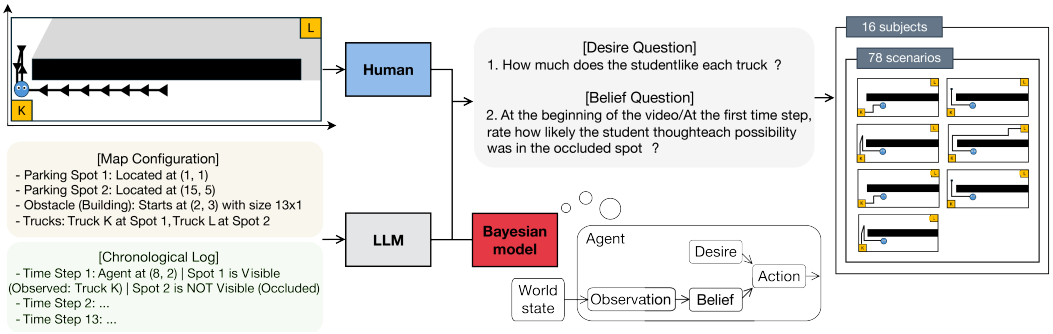


Figure 1: Task design. Sixteen human subjects and ten LLMs jointly inferred an agent's desire and belief from a spatiotemporal record of its actions. The top-left panel shows an example scenario where the agent navigates from the center to the left then upward to check the top-right food truck (L); gray shading marks the region out of agent's line-of-sight (isovist field). The 78 scenarios varied in the agent's path, food truck locations, and map configuration, and each was presented under End-step and Every-step conditions. Bayesian approach to ToM builds an internal model of agents relating world states, observations, beliefs, desire, and actions.

The food truck task was developed to study dynamic ToM in humans (Baker et al., 2017), whose judgments aligned closely with the joint desire-belief inference of the BToM model (desire correlation: $r = 0.91$, belief: $r = 0.78$). We test LLMs under two conditions. The End-step condition reproduces the original procedure, so that the archived human ratings and the BToM predictions serve as human and normative benchmarks, respectively. The Every-step condition, introduced here, holds the stimulus and questions fixed but asks for ratings after every step, using only the evidence up to that step (the full trajectory remains in the prompt). Collecting matching human data for this condition is not practical: each participant would have to report desire and belief after every step of every scenario, and such repeated querying within the same subject would itself alter later judgments. Bayesian models, in contrast, yield a posterior at every step and therefore serve as the normative benchmark for the Every-step condition.

The food truck task has a factorial design across 78 scenarios, where three food trucks—Korean (K), Lebanese (L), and Mexican (M)—could occupy either of two parking spots, Goal 1 (G1) at the bottom left and Goal 2 (G2) at the top right corner of a two-dimensional space (Figure 1, more details in Appendix B). Following Baker et al. (2017), truck K was fixed at G1 and scenarios varied only in whether G2 held truck L, truck M, or Nothing (N). We additionally fixed the labels, so M is never on campus and the belief question ranges over L, M, or N at G2. Importantly, a horizontal wall in the middle of the map blocks the agent's view of G2, so the agent must move toward G2 to observe what is there. Agents are assumed to be rational: each step has a small cost, but they will walk further for a preferred truck. Finally, the observer rates on a 7-point scale the agent's desire for each truck and the agent's initial belief about which truck (if any) is at G2.

Unlike the human participants, who viewed sequential images, LLMs received the spatial coordinates of the map and of the agent's movements, so that visual processing would not become a bottleneck on the measurement of ToM. For the same reason, the log states at each time step which parking spots were visible from the agent's position and what was observed there (Appendix A), rather than requiring the model to derive the line of sight. We avoided verbal descriptions of the scenes, which risk introducing pragmatic biases (Hu et al., 2025).

3.2 BToM Model

At the core of the BToM model are Partially Observable Markov Decision Processes (POMDPs) that formalize planning problems where an agent operates under imperfect knowledge of the world state. In the food truck task, the wall blocks the agent's line of sight to G2, so the agent has to walk toward it even though every step is costly. The agent therefore maintains a belief state—a distribution over possible world states—and updates it after each action and observation (Kaelbling et al., 1998):

$$b_t(s_t) \propto P(o_t|s_t) \sum_{s_{t-1}} P(s_t|s_{t-1}, a_{t-1}) b_{t-1}(s_{t-1}) \tag{1}$$

In Equation 1, the updated belief $b_t(s_t)$ combines the observation function $P(o_t|s_t)$, the transition probability $P(s_t|s_{t-1}, a_{t-1})$, and the prior belief $b_{t-1}(s_{t-1})$. Actions are selected by a softmax over belief-state action values with inverse temperature β. The BToM observer inverts this agent model to infer the agent's reward function and beliefs jointly. When estimating an earlier belief from the full trajectory, later steps enter only as a likelihood over what the agent did next; the world state the agent later observes is not evidence about that earlier belief, since the agent had no access to it at the time (Appendix D).

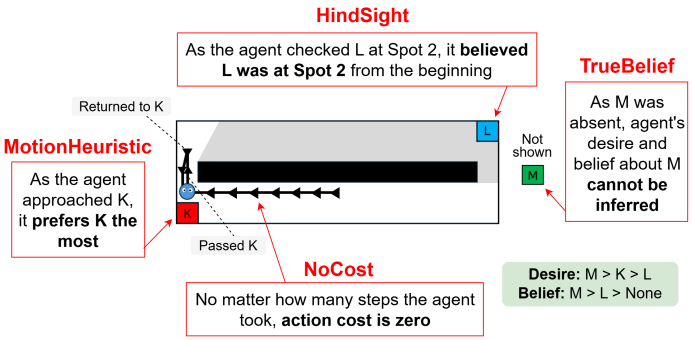


Figure 2: Variants of BToM: Four suboptimal alternative models. An example of "Check-GoBack" scenario is shown. The agent approaches the Korean truck (K, red) but passes it by, moves upward to check the food truck at top-right spot, and finally returns to K. Reasoning of the BToM model is as follows: the agent prefers Mexican (M) to Korean (K) and initially believed M to be at the top-right to explain agent's "check" behavior. K is preferred to Lebanese (L) as the agent goes back to eat K (green box for this ground truth).

To identify which components of BToM account for LLMs' behavior, we compared four alternative models, each corresponding to a misreading of the scenario in Figure 2: attending only to the final destination makes K the clear favorite (MotionHeuristic); projecting the agent's later knowledge back onto its initial belief makes L certain from the outset (Hindsight); discounting the steps spent on the detour removes the evidence that K was not strongly preferred (NoCost); and granting the agent full knowledge of the world state leaves nothing to infer about the unseen M (TrueBelief). The first three follow Baker et al. (2017); Hindsight is introduced here (Appendix D).

3.3 Analysis

We selected ten LLMs spanning four model families and both flagship and lightweight tiers: GPT-4o (Hurst et al., 2024), GPT-5.4 (Singh et al., 2025), GPT-o4-mini (OpenAI, 2025), Gemini-2.5-Flash, Gemini-2.5-Pro (Comanici et al., 2025), DeepSeek-V3.2 (Liu et al., 2025), DeepSeek-R1 (Guo et al.,

2025), DeepSeek-V4-Pro (Xu et al., 2026), Claude-4.6-Sonnet (Anthropic, 2026b), and Claude-4.6-Opus (Anthropic, 2026a). All models were queried in their default configuration, so our claims concern default deployment behavior; because several of these models reason by default whereas others do not, reasoning mode is not controlled across models (Appendix C). In the Every-step condition, GPT-4o and DeepSeek-V3.2 returned malformed stepwise outputs in most runs and were excluded.

Both End-step and Every-step conditions provide the full trajectory via a single prompt, listing for each time step the agent's movement coordinates and which trucks at G1 and G2 were visible or not from that position. Models rated on a 7-point scale their desire for trucks K, L, and M and their belief that L, M, or N was at G2 (see Appendix A for full prompts). The three belief ratings of each scenario were normalized to sum to one so that they could be compared with the model posteriors, whereas desire ratings were compared on their original scale (Baker et al., 2017). Each of 78 scenarios was run 16 times per model through independent API calls, and ratings were averaged over runs. As a similarity metric between humans, BToM, and LLMs, we computed Pearson correlations of their ratings, pooled over scenarios and options, separately for desire and belief; 95% confidence intervals were obtained from 2,000 bootstrap resamples of scenarios, with the three options of a scenario resampled together. We further performed a variance partitioning analysis to identify which of BToM, TrueBelief, NoCost, MotionHeuristic, and Hindsight uniquely accounted for each LLM's rating variance: the unique variance of a model is the decrease in adjusted R^2 when its predictions are removed from a linear regression that includes all five models. Five irrational scenarios were excluded from both analyses (Baker et al., 2017). Consequently, the final dataset comprises 219 data points for each desire and belief query (73 scenarios $\times$ 3 options).

The inverse temperature β was fitted to the human data for each observer model that has one, consistent with the original work (Baker et al., 2017); Hindsight's additional parameter is described in Appendix D. Both conditions were analyzed with correlation and variance partitioning.

4 RESULTS

We report how closely LLMs approximated the human and BToM benchmarks in the End-step condition (§4.1), which alternative model best accounts for the lower-performing models (§4.2), and how the better-performing models behave when judgments are elicited at every step (§4.3).

4.1 LLM ALIGNMENT WITH HUMANS AND BTOM

In the End-step condition, LLMs received the complete trajectory followed by a single set of desire and belief questions, as human participants did. We correlated each LLM's ratings of the three options (K, L, M for desire; L, M, N for belief) with those of humans, BToM, and the alternative models. For the diagnostic analyses, we labeled an LLM "Better" if its correlation with humans exceeded its correlation with every alternative model in desire or belief, and "Worse" otherwise; this is a descriptive stratification, not a ranking.

Figure 3 shows representative Better (Claude-4.6-Opus) and Worse (GPT-5.4) models. As in Baker et al. (2017), BToM reproduced the human desire and belief profiles (Figure 3a left); Claude-4.6-Opus produced a similar profile, whereas GPT-5.4 departed from it (Figure 3a right). Pooled over scenarios and options, GPT-5.4 correlated weakly with human and BToM ratings (negatively with BToM for desire), whereas Claude-4.6-Opus correlated with both at values close to the human–BToM alignment.

Applying this analysis to all ten LLMs, six fell into the Worse group and four into the Better group (Figure 6; all values in Table 4, Appendix E). The Better models were Gemini-2.5-Pro ($r_{desire} = 0.78$, [0.70, 0.83] for 95% confidence interval (CI); $r_{belief} = 0.90$, [0.86, 0.93]), DeepSeek-R1 (0.81, [0.75, 0.86]; 0.59, [0.49, 0.69]), DeepSeek-V4-Pro (0.84, [0.79, 0.88]; 0.73, [0.64, 0.80]), and Claude-4.6-Opus (0.86, [0.81, 0.91]; 0.76, [0.61, 0.86]). At the other extreme, the correlations for GPT-4o and GPT-5.4 were not significantly different from zero in both dimensions, and DeepSeek-V3.2's belief correlation was significantly below zero showing clear misalignment to human data (-0.28, $[-0.40, -0.16]$).

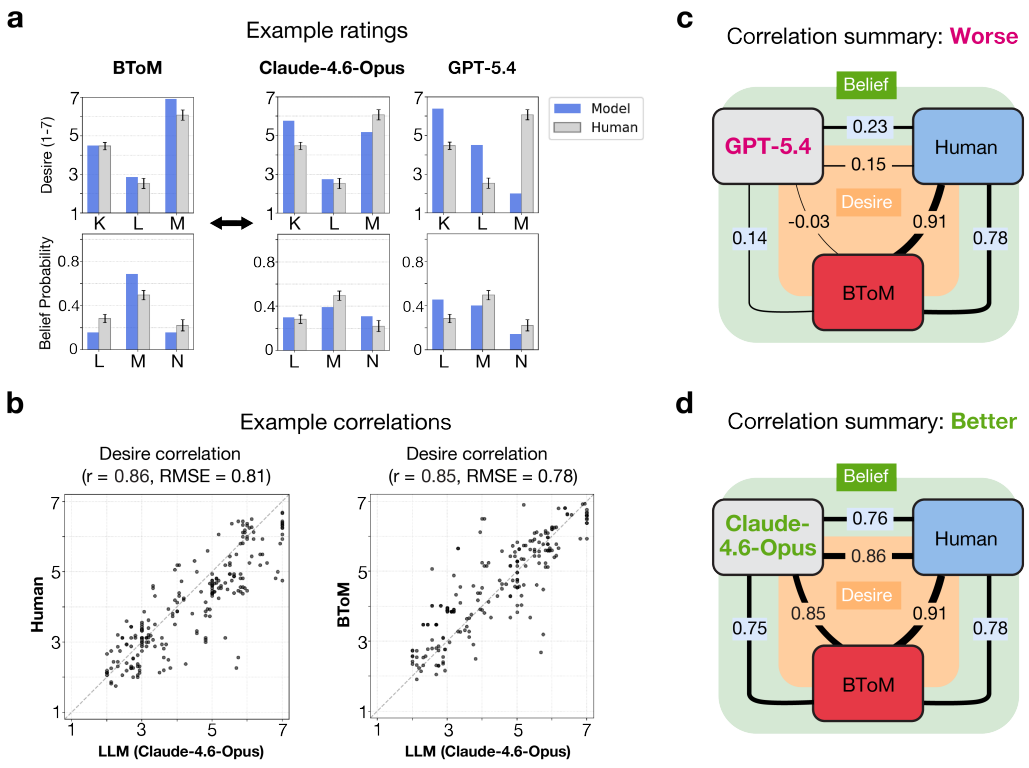


Figure 3: LLM-BToM-human comparison in the End-step condition. (a) Raw ratings of BToM, Claude-4.6-Opus and GPT-5.4 are shown with human benchmark (top row for desire; bottom row for belief). (b) Correlations of all desire ratings in Claude-4.6-Opus with human (left) and BToM (right) ratings are shown. (c) Summary of the rating correlations of GPT-5.4 is shown. Overall, the correlations were lower than human-BToM correlations both for belief and desire. (d) Rating correlations of Claude-4.6-Opus to human or BToM are comparable to human-BToM correlation, both in desire and belief.

No LLM reached the human–BToM alignment ($r_{desire} = 0.91$; $r_{belief} = 0.78$) in both dimensions. Notably, Gemini-2.5-Pro's belief ratings correlated with human ratings more strongly (CI: [0.86, 0.93]) than BToM does (CI: [0.70, 0.85]), suggesting that it captures systematic human deviations from the normative model. Intervals within the Better group overlap substantially (e.g., Claude-4.6-Opus and DeepSeek-R1 in desire), so we use the grouping only to stratify the diagnostic analyses below. Claude-4.6-Opus had the highest desire correlation and was the only LLM best explained by BToM in both dimensions (Section 4.3), while Gemini-2.5-Pro was closest to humans in belief. Note that the human–BToM values are the cross-validated correlations reported by Baker et al. (2017), so this comparison is approximate.

4.2 Cost-insensitive desire inference

To identify what separates the worse-performing LLMs from human data, we compared their ratings with the four alternative observer models, each of which removes or distorts a critical component of BToM (Section 3.2). For desire, the pattern was uniform: every Worse model correlated most strongly with NoCost. For belief, no single alternative dominated. The alternative models with the highest correlation were TrueBelief (GPT-4o), NoCost (Gemini-2.5-Flash, DeepSeek-V3.2), MotionHeuristic (GPT-o4-mini, Claude-4.6-Sonnet), and Hindsight (GPT-5.4). Pooling the Worse models and recomputing correlations at the group level confirmed this asymmetry (Figure 4a): NoCost again explained desire best ($r = 0.80$), whereas TrueBelief ($r = 0.59$), NoCost ($r = 0.59$), and Hindsight ($r = 0.61$) accounted for belief similarly; the belief result therefore does not single out any one account.

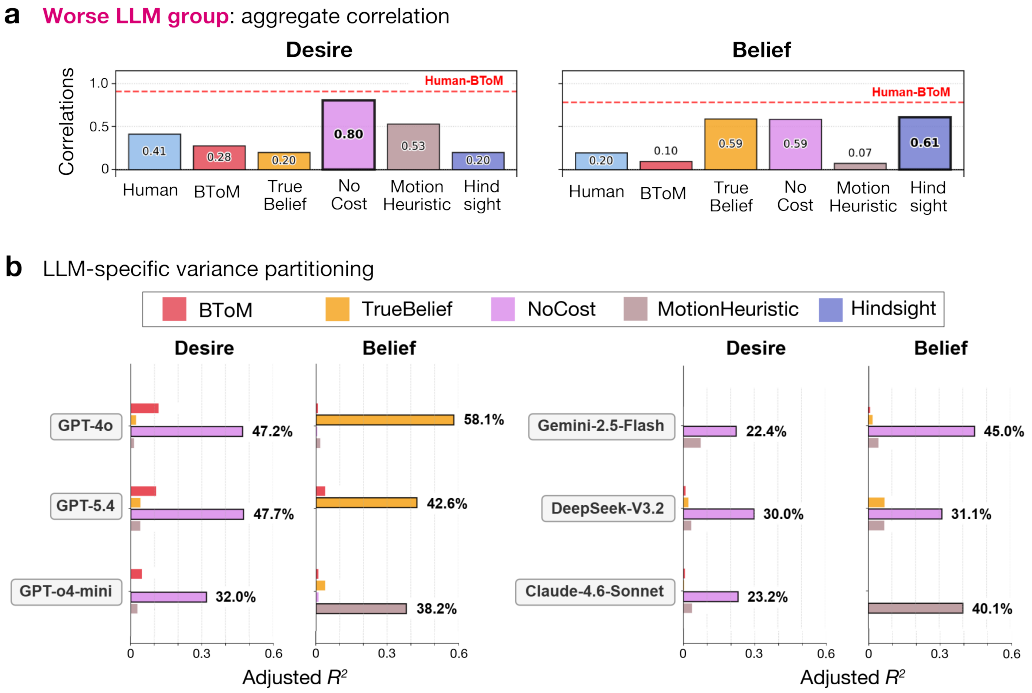


Figure 4: Ablation analysis of worse-performing LLMs. (a) Correlations on data pooled over the six Worse models: desire ratings are best captured by NoCost ($r = 0.80$), whereas TrueBelief, NoCost, and Hindsight account for belief ratings about equally ($r = 0.59$–0.61). (b) Variance partitioning analysis points out the LLM-specific best suboptimal models. For instance, GPT-4o was best captured by the NoCost model for desire and the TrueBelief model for belief (same color legend as panel a). Per-model correlations are in Figure 7.

We then applied variance partitioning to isolate the rating variance uniquely attributable to each model (Figure 4b), which mitigates the multicollinearity between BToM and its alternatives that correlation alone cannot resolve. NoCost dominated desire in every Worse model, whereas the dominant source of belief variance differed across models. Because unique variance and zero-order correlation can disagree when predictors are collinear, the best belief account for an individual model can differ between the two analyses (e.g., GPT-5.4: Hindsight by correlation, TrueBelief by variance partitioning). Taken together, the desire judgments of the worse-performing LLMs resemble those of an observer who ignores the agent's action costs, while their belief judgments follow idiosyncratic strategies.

4.3 Hindsight Bias in Stepwise Belief

Applying the same variance partitioning to the better-performing LLMs in the End-step condition (Appendix F), the unique variance explained by BToM increased for desire relative to the Worse group, whereas the best account of belief varied across LLMs; only Claude-4.6-Opus was best explained by BToM in both dimensions (unique adjusted R^2: 0.174 for desire, 0.135 for belief; small by construction because BToM and its alternatives make highly correlated predictions).

BToM also specifies how the inference should evolve step by step: its reference at step t is the posterior given steps 1 to t, exactly the evidence the Every-step instructions allow. We compared the Every-step ratings of the four better-performing models with this normative benchmark.

In a representative scenario (Figure 5a), BToM revised its estimates at two critical steps: "Pass G1" (K less preferred) and "Return G1" (K preferred to L; M initially believed to be at G2). Claude-4.6-

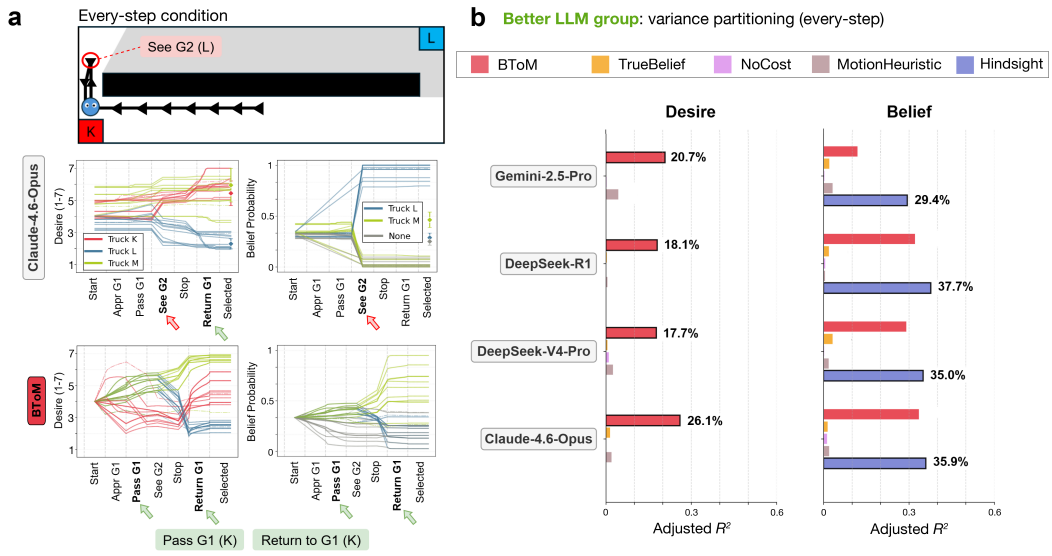


Figure 5: Stepwise responses in Every-step condition. (a) In the same "Check-GoBack" scenario as Figure 2, the desire and belief ratings of Claude-4.6-Opus (middle row) are shown as a function of individual action steps. At bottom row, BToM's inference trajectory is shown as a normative baseline. Each line denotes one truck option, and the diamonds mark the corresponding End-step ratings. The BToM curve at step t uses only steps 1 to t, the same evidence the LLM was instructed to use. Red and green arrows highlight key phases where the rating scores exhibit abrupt fluctuations differently between LLM and BToM model. (b) Variance partitioning analysis of better-performing LLMs' stepwise ratings shows that while inferred desires can be well explained by BToM (left column), the inferred belief is best captured by the Hindsight model.

Opus instead shifted its belief sharply at "See G2", when the agent sees which truck is at G2. That step is decisive for the agent's *current* belief but should not revise its *initial* belief.

Pooled over the four models, final Every-step ratings—made with the same information as End-step—agreed with End-step ratings for desire but not for belief ($r_{desire} = 0.91$, $r_{belief} = 0.02$), and final Every-step belief ratings no longer correlated with human ratings ($r_{desire} = 0.82$, $r_{belief} = -0.02$; End-step, same pooling: 0.85, 0.82).

Variance partitioning of the Every-step data (Figure 5b) showed that BToM still accounted for the largest unique variance in desire, whereas Hindsight did so for belief in all four LLMs. That is, the stepwise belief judgments treated the later-observed world state as evidence about the agent's earlier belief, whereas a normative observer uses only the agent's later *actions* (Appendix D). Reasoning traces from Claude-4.6-Opus show this pattern explicitly: because L was later seen at G2, the agent "would now believe L was at Spot 2 with certainty" at $t = 1$ (Appendix G). These findings suggest that belief, unlike desire, was not stable across elicitation formats, a prompt-dependent departure from the normative model consistent with hindsight bias.

5 Discussion and Conclusion

We evaluated the ToM of LLMs in a dynamic spatial task against human ratings and a normative Bayesian model. The best LLMs approached the human–BToM alignment without exceeding it in both dimensions. Lesioned BToM models showed that the desire judgments of worse-performing LLMs resemble an observer who ignores action costs, whereas their belief judgments follow heterogeneous suboptimal accounts; and when better-performing LLMs gave stepwise judgments, their beliefs were best captured by the Hindsight model. Across analyses, desire judgments were robust and belief judgments fragile: desire alignment was preserved across models and elicitation formats, whereas belief alignment varied across models and vanished in the Every-step condition. Because

both were inferred from the same trajectories, this dissociation reflects the mental state being inferred rather than the stimuli, and it parallels, without implying a common mechanism, the developmental ordering in which children understand desires before beliefs (Wellman & Liu, 2004).

The main contribution of this work lies in explaining the behavioral differences between LLMs with an explicit computational model of human social reasoning (Baker et al., 2017). Prior dynamic ToM work has shown that LLMs struggle to retrieve earlier belief states over extended interactions (Nguyen et al., 2026) and that their mental-state inferences can be inconsistent with their behavior predictions (Muchovej et al., 2026). We go further by comparing competing inverse-planning explanations of such departures, which identified cost-insensitive desire inference (NoCost) and hindsight-contaminated retrospective belief inference (Hindsight) as specific candidate accounts. A conventional accuracy-based ToM benchmark scores each response against a single correct answer; the tripartite comparison among human judgments, normative BToM predictions, and LLM outputs instead localizes where a model departs from human-like inference, by identifying which component of BToM—the weighing of action costs, belief updating, or the separation of the agent's knowledge from the observer's—its judgments violate. The model also supplies a benchmark where human data are difficult to obtain: BToM yields a normative posterior at every step of the trajectory, whereas eliciting stepwise judgments from human participants would require repeated within-subject querying that would itself alter later judgments. Evaluating the Every-step condition against this benchmark revealed that models whose end-of-trajectory judgments closely approximated human ratings nonetheless produced markedly different initial-belief estimates when asked for successive judgments, which identifies a concrete target for future work on the robustness of LLM social inference.

The desire judgments of weaker LLMs were best described by an observer that ignores the cost of the agent's actions. Human desire attribution rests on a cost-benefit reading of action (Jara-Ettinger et al., 2016): choosing a longer path to a truck implies a stronger desire for it. Similarity to NoCost is consistent with insensitivity to this trade-off, but could also arise if a model attends mainly to where the agent ends up or assumes a different effort scale; manipulating path length while holding outcome fixed would separate these accounts. Cost insensitivity would limit LLMs in embodied settings, where the motivation behind an action matters as much as its outcome.

Robust ToM requires dissociating objective facts from an agent's beliefs. Given the full trajectory and a single question, the better-performing models did so at close to human levels, but when queried at every step with the full trajectory in context, their judgments drifted toward the state the agent later observed, consistently across all four models. Humans show related biases: outcome knowledge distorts reconstructions of one's own earlier judgments (hindsight bias; Fischhoff, 1975), and people attribute their own knowledge to less informed others (curse of knowledge; Camerer et al., 1989; Birch & Bloom, 2007). Our pattern is closer to the latter, since the observer's later knowledge is attributed to the agent's earlier belief. We speculate that attention over the entire log, which makes retrospective inference possible, also lets the attributed epistemic state become omniscient across time points; multi-turn prompting in which models never see future steps (Zhang et al., 2024) would test this. A comparison of the Every-step format with Chain-of-Thought prompting is given in Appendix H.

Our work has several limitations that warrant further investigation. The Every-step condition differs from End-step not only in stepwise elicitation but also in output length, bookkeeping, and the demand to ignore in-context future steps, and its belief question can be read as asking about the agent's current belief; online multi-turn and prefix-truncated conditions would separate these factors. Human data exist only for End-step, so the Every-step comparison relies on the normative benchmark, and Hindsight was introduced to describe the Every-step pattern (Appendix D). LLMs received coordinates rather than images, and we did not independently verify their parsing of the spatial log. We used a single prompt wording and fixed truck labels, so memorization and label biases cannot be excluded, and models were queried in default configurations that differ in reasoning mode. Finally, the stimuli are two-dimensional with a single occluding wall; naturalistic stimuli (Jia et al., 2026) and vision-language models are natural next steps. BToM's causal structure over state, observation, belief, desire, and action can be selectively ablated, and further alternative models can extend this diagnostic toolkit.

REFERENCES

Anthropic. System card: Claude opus 4.6. *https://www.anthropic.com/claude-opus-4-6-system-card/*, 2026a.

Anthropic. System card: Claude sonnet 4.6. *https://www.anthropic.com/claude-sonnet-4-6-system-card/*, 2026b.

Chris L Baker, Julian Jara-Ettinger, Rebecca Saxe, and Joshua B Tenenbaum. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1(4): 0064, 2017.

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

Arghavan Bazigaran and Hansem Sohn. Concept generalization in humans and large language models: Insights from the number game. *arXiv preprint arXiv:2512.20162*, 2025.

Susan A. J. Birch and Paul Bloom. The curse of knowledge in reasoning about false beliefs. *Psychological Science*, 18(5):382–386, 2007.

Paul Bloom and Tim P German. Two reasons to abandon the false belief task as a test of theory of mind. *Cognition*, 77(1):B25–B31, 2000.

Matteo Bortoletto, Constantin Ruhdorfer, and Andreas Bulling. Tom-ssi: Evaluating theory of mind in situated social interactions. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 32252–32277, 2025.

Colin Camerer, George Loewenstein, and Martin Weber. The curse of knowledge in economic settings: An experimental analysis. *Journal of Political Economy*, 97(5):1232–1254, 1989.

Peter Carruthers. Two systems for mindreading? *Review of Philosophy and Psychology*, 7(1): 141–162, 2016.

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

Nicole Drummond and Yael Niv. Model-based decision making and model-free learning. *Current Biology*, 30(15):R860–R865, 2020.

Baruch Fischhoff. Hindsight $\neq$ foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance*, 1(3): 288–299, 1975.

Kim Frost. On the very idea of direction of fit. *Philosophical Review*, 123(4):429–484, 2014.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D. Goodman. Understanding social reasoning in language models with language models. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, volume 36, 2023.

Rebekah A Gelpí, Eric Xue, and William A Cunningham. Towards machine theory of mind with large language model-augmented inverse planning. *arXiv preprint arXiv:2507.03682*, 2025.

Michael Georgeff, Barney Pell, Martha Pollack, Milind Tambe, and Michael Wooldridge. The belief-desire-intention model of agency. In *International workshop on agent theories, architectures, and languages*, pp. 1–10. Springer, 1998.

Mariel K Goddu and Alison Gopnik. The development of human causal learning and reasoning. *Nature Reviews Psychology*, 3(5):319–339, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Jessica B Hamrick. Analogues of mental simulation and imagination in deep learning. *Current Opinion in Behavioral Sciences*, 29:8–16, 2019.

Jennifer Hu, Felix Sosa, and Tomer Ullman. Re-evaluating theory of mind evaluation in large language models. *Philosophical Transactions B*, 380(1932):20230499, 2025.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Julian Jara-Ettinger, Hyowon Gweon, Laura E. Schulz, and Joshua B. Tenenbaum. The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences*, 20(8):589–604, 2016.

Jingru Jia, Zehua Yuan, Junhao Pan, Paul McNamara, and Deming Chen. Llm strategic reasoning: Agentic study through behavioral game theory. *Advances in Neural Information Processing Systems*, 38:58318–58350, 2026.

Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer Ullman, Antonio Torralba, Joshua Tenenbaum, and Tianmin Shu. Mmtom-qa: Multimodal theory of mind question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16077–16102, 2024.

Nitish Joshi, Abulhair Saparov, Yixin Wang, and He He. Llms are prone to fallacies in causal inference. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 10553–10569, 2024.

Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

Subbarao Kambhampati, Karthik Valmeekam, Lin Guan, Mudit Verma, Kaya Stechly, Siddhant Bhambri, Lucas Paul Saldyt, and Anil B Murthy. Position: Llms can't plan, but can help planning in llm-modulo frameworks. In *Forty-first International Conference on Machine Learning*, 2024.

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.

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

Alexander Y Ku, Declan Campbell, Xuechunzi Bai, Jiayi Geng, Ryan Liu, Raja Marjieh, R Thomas McCoy, Andrew Nam, Ilia Sucholutsky, Veniamin Veselovsky, et al. Levels of analysis for large language models. *arXiv preprint arXiv:2503.13401*, 2025.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*, 2025.

David Liu, Andrew N Meltzoff, and Henry M Wellman. Neural correlates of belief-and desire-reasoning. *Child development*, 80(4):1163–1171, 2009.

Ryan Liu, Jiayi Geng, Joshua C Peterson, Ilia Sucholutsky, and Thomas L Griffiths. Large language models assume people are more rational than we really are. *arXiv preprint arXiv:2406.17055*, 2024.

Jiahao Lu, Ziwei Xu, and Mohan Kankanhalli. Reasoning llms are wandering solution explorers. *arXiv preprint arXiv:2505.20296*, 2025.

Shima Rahimi Moghaddam and Christopher J Honey. Boosting theory-of-mind performance in large language models via prompting. *arXiv preprint arXiv:2304.11490*, 2023.

John Muchovej, Amanda Royka, Shane Lee, and Julian Jara-Ettinger. Gpt-4o lacks core features of theory of mind. *arXiv preprint arXiv:2602.12150*, 2026.

Thuy Ngoc Nguyen, Duy Nhat Phan, and Cleotilde Gonzalez. Dynamic theory of mind as a temporal memory problem: Evidence from large language models. *arXiv preprint arXiv:2603.14646*, 2026.

OpenAI. Openai o3 and o4-mini system card. *https://openai.com/index/o3-o4-mini-system-card/*, 2025.

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

Linlu Qiu, Fei Sha, Kelsey Allen, Yoon Kim, Tal Linzen, and Sjoerd van Steenkiste. Bayesian teaching enables probabilistic reasoning in large language models. *Nature Communications*, 2026.

Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, SM Ali Eslami, and Matthew Botvinick. Machine theory of mind. In *International conference on machine learning*, pp. 4218–4227. PMLR, 2018.

Betty M Repacholi and Alison Gopnik. Early reasoning about desires: evidence from 14-and 18-month-olds. *Developmental psychology*, 33(1):12, 1997.

Moein Salimi, Shaygan Adim, Danial Parnian, Nima Alighardashi, Mahdi Jafari Siavoshani, and Mohammad Hossein Rohban. Wiring the'why': A unified taxonomy and survey of abductive reasoning in llms. *arXiv preprint arXiv:2604.08016*, 2026.

Rebecca Saxe and Sean Dae Houlihan. Formalizing emotion concepts within a bayesian model of theory of mind. *Current opinion in Psychology*, 17:15–21, 2017.

Rebecca Saxe and Nancy Kanwisher. People thinking about thinking people: the role of the temporo-parietal junction in "theory of mind". In *Social neuroscience*, pp. 171–182. Psychology Press, 2013.

Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. Minding language models' (lack of) theory of mind: A plug-and-play multi-character belief tracker. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 13960–13980, 2023.

Rose M Scott and Renée Baillargeon. Early false-belief understanding. *Trends in cognitive sciences*, 21(4):237–249, 2017.

John R Searle. *Intentionality: An essay in the philosophy of mind*. Cambridge university press, 1983.

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 *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 2257–2273, 2024.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Hansem Sohn and Mehrdad Jazayeri. Validating model-based bayesian integration using prior–cost metamers. *Proceedings of the National Academy of Sciences*, 118(25):e2021531118, 2021.

Peiyang Song, Pengrui Han, and Noah Goodman. Large language model reasoning failures. *arXiv preprint arXiv:2602.06176*, 2026.

James WA Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, et al. Testing theory of mind in large language models and humans. *Nature human behaviour*, 8(7):1285–1295, 2024.

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

Tomer D Ullman and Joshua B Tenenbaum. Bayesian models of conceptual development: Learning as building models of the world. *Annual Review of Developmental Psychology*, 2(1):533–558, 2020.

Zehong Wang, Fang Wu, Hongru Wang, Xiangru Tang, Bolian Li, Zhenfei Yin, Yijun Ma, Yiyang Li, Weixiang Sun, Xiusi Chen, et al. Why reasoning fails to plan: A planning-centric analysis of long-horizon decision making in llm agents. *arXiv preprint arXiv:2601.22311*, 2026.

Zihan Wang, Isaac Davis, and Julian Jara-Ettinger. Modeling other minds: A computational account of social cognition and its development. *Annual Review of Developmental Psychology*, 7, 2025.

Henry M Wellman and David Liu. Scaling of theory-of-mind tasks. *Child development*, 75(2): 523–541, 2004.

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

Yuheng Wu, Jianwen Xie, Denghui Zhang, and Zhaozhuo Xu. Del-tom: Inference-time scaling for theory-of-mind reasoning via dynamic epistemic logic. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 11394–11408, 2025.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

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

Ilker Yildirim and LA Paul. From task structures to world models: what do llms know? *Trends in Cognitive Sciences*, 28(5):404–415, 2024.

Lance Ying, Tan Zhi-Xuan, Lionel Wong, Vikash Mansinghka, and Joshua Tenenbaum. Grounding language about belief in a bayesian theory-of-mind. *arXiv preprint arXiv:2402.10416*, 2024.

Lance Ying, Tan Zhi-Xuan, Lionel Wong, Vikash Mansinghka, and Joshua B Tenenbaum. Understanding epistemic language with a language-augmented bayesian theory of mind. *Transactions of the Association for Computational Linguistics*, 13:613–637, 2025.

Fangxu Yu, Lai Jiang, Shenyi Huang, Zhen Wu, and Xinyu Dai. Persuasivetom: A benchmark for evaluating machine theory of mind in persuasive dialogues. *arXiv preprint arXiv:2502.21017*, 2025.

Zheng Zhang, Hossein Amiri, Zhenke Liu, Liang Zhao, and Andreas Züfle. Large language models for spatial trajectory patterns mining. In *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Geospatial Anomaly Detection*, pp. 52–55, 2024.

Zhining Zhang, Chuanyang Jin, Mung Yao Jia, and Tianmin Shu. Autotom: Automated bayesian inverse planning and model discovery for open-ended theory of mind. In *ICLR 2025 Workshop on Foundation Models in the Wild*, 2025.

APPENDIX

The Appendix is organized as follows: we elaborate on experiment methods including full prompts (A), task details (B), model specifications (C) and BToM model (D). The following supplementary results consist of the correlation analysis with per-model statistics (E), the variance partitioning analysis (F), the inspection of reasoning logs (G), and a comparison of Every-step prompting with Chain-of-Thought (H).

SUPPLEMENTARY METHODS

A FULL PROMPTS

Here, we provide the full prompts used to test the food truck task for LLMs. The user prompt shown is for the first scenario as a case example, yet in the experiment, the map configuration and the chronological log always changed according to each scenario's settings. Some of the wording (e.g., experiment instructions, questions) was slightly modified from the human experiment (Baker et al., 2017) because we changed the input stimuli from images to texts with coordinates (see Methods for rationale): therefore, we attempted to provide more task context in the prompt and particularly emphasized the retrospective belief inference task—to retroactively infer the agent's initial belief about the specific truck being at G2—by using terms such as "At the first time step" and "at t=1". The system prompt retains phrases from the human instructions that refer to visual elements (face icons, black and yellow squares, grey and white areas); LLMs received no images, and the corresponding information was conveyed by the coordinates and visibility annotations in the log.

The user prompt and the system prompt for the End-step condition were given as the following:

End-step: System Prompt

You are a participant in a psychology experiment designed to infer intentions from behavior. Please make judgments based strictly on the provided setting descriptions.

[Experiment Settings]
1. **Context**: Students (represented as face icons) at the Manitoba Institute of Technology (MIT) love food trucks. They already know the locations of the buildings (black squares) and the parking spots (yellow squares).
2. **Trucks**: Three different trucks (Mexican = M, Korean = K, Lebanese = L) compete for parking spots daily. There are 2 parking spots in total. Spots can be occupied by a truck or be empty.
3. **Visibility**: Students CANNOT see the 'grey area' occluded by the building (Invisible). They CAN see the 'white area' (Visible). For parking spots they cannot currently see, students rely on prior experience to know, believe, or guess which truck is there.
4. **Motivation**: Students are very hungry and want to eat immediately, but they are willing to walk further if there is a food truck they prefer more.

The logs provided represent the decision-making process of different students on different days as they move around the campus to decide where to have lunch.

[Output Format]
Please provide the response in JSON format. Do not include any markdown formatting like "'json.

End-step: User Prompt

Now, please analyze the scenario provided below.

[Map Configuration]
- Parking Spot 1: Located at (1, 1)
- Parking Spot 2: Located at (15, 5)
- Obstacle (Building): Starts at (2, 3) with size 13x1
- Trucks: Truck K at Spot 1, Truck L at Spot 2

[Chronological Log]
Time Step 1: Agent at (2, 2) — Spot 1 is Visible (Observed: Truck K) — Spot 2 is NOT Visible (Occluded)
Time Step 2: Agent at (1, 2) — Spot 1 is Visible (Observed: Truck K) — Spot 2 is NOT Visible (Occluded)

Time Step 3: Agent at (1, 3) — Spot 1 is Visible (Observed: Truck K) — Spot 2 is NOT Visible (Occluded)
Time Step 4: Agent at (1, 4) — Spot 1 is Visible (Observed: Truck K) — Spot 2 is Visible (Observed: Truck L)
Time Step 5: Agent at (1, 4) — Spot 1 is Visible (Observed: Truck K) — Spot 2 is Visible (Observed: Truck L)
Time Step 6: Agent at (1, 3) — Spot 1 is Visible (Observed: Truck K) — Spot 2 is NOT Visible (Occluded)
Time Step 7: Agent at (1, 2) — Spot 1 is Visible (Observed: Truck K) — Spot 2 is NOT Visible (Occluded)
Time Step 8: Agent at (1, 1) — Spot 1 is Visible (Observed: Truck K) — Spot 2 is NOT Visible (Occluded)

Based on the chronological situation above, please perform the following tasks:

1. At the LAST time step, rate the student's preference for Truck K, L, and M. (Scale: 1 = Dislike strongly to 7 = Like strongly)

2. At the FIRST time step, rate the student's likelihood for L, M, and Empty being in the occluded spot at t=1. (Scale: 1 = Definitely not there to 7 = Definitely there)

In contrast to the End-step condition, we instructed forward sequential processing in the Every-step condition, prohibiting the use of information from steps ahead of time point t when belief and desire states at t were queried. Because the full trajectory remains in the prompt, this instruction cannot be enforced; the condition asks whether LLMs can withhold knowledge that is present in their context, and use only the accumulated evidence up to t from the agent's perspective. Note that the human subjects only completed the End-step condition as it is not practical to ask individual subjects to report inferred belief and desire at every step across many task conditions. We can, however, test BToM model and LLMs for this Every-step condition without the practical limitation.

The user prompt and the system prompt for the Every-step condition were given as the following (for the user prompt, the map configuration and chronological log were identical to the End-step prompt and are omitted here):

Every-step: System Prompt

You are a participant in a psychology experiment designed to infer intentions from behavior. Please make judgments based strictly on the provided setting descriptions.

[Experiment Settings]
1. **Context**: Students (represented as face icons) at the Manitoba Institute of Technology (MIT) love food trucks. They already know the locations of the buildings (black squares) and the parking spots (yellow squares).
2. **Trucks**: Three different trucks (Mexican = M, Korean = K, Lebanese = L) compete for parking spots daily. There are 2 parking spots in total. Spots can be occupied by a truck or be empty.
3. **Visibility**: Students CANNOT see the 'grey area' occluded by the building (Invisible). They CAN see the 'white area' (Visible). For parking spots they cannot currently see, students rely on prior experience to know, believe, or guess which truck is there.
4. **Motivation**: Students are very hungry and want to eat immediately, but they are willing to walk further if there is a food truck they prefer more.
5. **Time step**: Crucially, agents only know what they have seen SO FAR.

The logs provided represent the decision-making process of different students on different days as they move around the campus to decide where to have lunch.

[STRICT RULES]
1. Sequential Processing: You must process the log strictly from Time Step 1 to the end.

2. NO Look-ahead: When analyzing Time Step 't', you must NOT use any information from future steps (t+1, t+2...).
- Pretend you are watching a live video feed and do not know the ending.
3. Update: Update the estimates ONLY based on the accumulated evidence up to the current step.
4. YES Look-back: When you're doing the belief analysis, you must retrospectively consider the agent's initial belief at "time_step": 1.
5. NO LAZINESS (CRITICAL): You must output data for EVERY SINGLE time step present in the log. Do NOT skip, abbreviate, or use placeholders like "...".

[Output Format]
Please provide the response in JSON format. Do not include any markdown formatting like "'json.

Every-step: User Prompt

Based on the chronological situation above, please perform the following tasks:

[CRITICAL REQUIREMENT]
This log has exactly 8 time steps. Your JSON array MUST contain exactly 8 objects. Do NOT use "..." or skip any steps.

1. At EVERY time step, rate the student's preference for Truck K, L, and M. (Scale: 1 = Dislike strongly to 7 = Like strongly)

2. At EVERY time step, rate the student's likelihood for L, M, and Empty being in the occluded spot at t=1 only given the current information. (Scale: 1 = Definitely not there to 7 = Definitely there)

Path group	G2 presence	Agent position	Environment	Path number
NoCheck	1) Yes 2) No	1) Left 2) Middle 3) Right	1) C-shape 2) O-shape 3) Backward C-shape	G2 present (n=9): 5, 10, 15, 42, 45, 48, 27, 30, 33 G2 absent (n=9): 18, 21, 24, 50, 52, 54, 35, 37, 39
Check-Partial				G2 present (n=12): 55, 63, **71**, 57, 65, 73, 59, 67, 75, 61, 69, 77 G2 absent (n=12): 56, 64, **72**, 58, 66, 74, 60, 68, 76, 62, 70, 78
Check-GoBack				G2 present (n=12): 1, 6, **11**, 40, 43, 46, 3, 8, 13, 25, 28, 31 G2 absent (n=12): 16, 19, **22**, 49, 51, 53, 17, 20, 23, 34, 36, 38
Check-Stay	Yes			G2 present (n=12): 2, 7, **12**, 41, 44, 47, 4, 9, 14, 26, 29, 32

Table 1: Factorial path design for the food truck paradigm. Numbers in **bold** indicate the irrational paths (11, 12, 22, 71, 72).

B Task Details

The 78 scenarios from Baker et al. are categorized into seven different groups based on the combination of environment and agent behavior: NoCheck (G2 present), Check-Partial (G2 present), Check-GoBack (G2 present), Check-Stay (G2 present), NoCheck (G2 absent), Check-Partial (G2 absent), and Check-GoBack (G2 absent), with five scenarios of irrational paths (Table 1). These categories differ in the food-truck state for Goal 2 (absent vs. present), the agent's initial (Check vs. NoCheck) and final (Stay, GoBack, Partial) movements toward G2.

In our experiment, NoCheck, Check-Partial, and irrational paths represent trivial ToM inferences, such as failing or rejecting to check G2, stopping without making a decision after checking, or taking a suboptimal detour. On the other hand, the core scenarios are Check-GoBack and Check-Stay, where the agent makes a decision to go back or stay after taking a long walk around the building to see the truck at G2. Among these, Check-GoBack is particularly suited for evaluating false belief understanding as the GoBack behavior reflects dynamic changes in agent's belief. Therefore, we focused on explaining Check-GoBack (G2 present) scenarios as a representative in the main text. In Baker et al., extensive randomizations (e.g., scenario order, truck labels, agent colors and names, visual layouts) were used to prevent within-subject confounds. However, our independent API initializations obviated the need for this randomization in LLMs.

C LLM Specifications

We queried each model through its provider's API. Table 2 lists the model identifiers, snapshots, and access dates. Temperature was left at the provider-fixed value of 1.0 for models that do not expose it and set to 0.7 otherwise; no other decoding parameters were changed. All models were evaluated in their default configuration. For models in which extended reasoning is optional and off by default (e.g., the Claude models), it was not enabled; models that reason by default were run with their default reasoning behavior. Reasoning mode therefore differs across models and is not a controlled factor in our comparisons. The reasoning traces in Appendix G were collected separately, from Claude-4.6-Opus with extended reasoning enabled, in five additional runs on the Check-GoBack scenarios; for Claude 4 models, the API returns a summarized version of the reasoning rather than the full trace.

Model	Version
GPT-4o	gpt-4o-2024-08-06
GPT-5.4	gpt-5.4-2026-03-05
GPT-o4-mini	o4-mini-2025-04-16
Gemini-2.5-Flash	gemini-2.5-flash (February 2026)
Gemini-2.5-Pro	gemini-2.5-pro (March, May 2026)
DeepSeek-V3.2	deepseek-chat (March 2026)
DeepSeek-R1	deepseek-reasoner (March 2026)
DeepSeek-V4-pro	deepseek-v4-pro (June 2026)
Claude-4.6-Sonnet	claude-sonnet-4-6
Claude-4.6-Opus	claude-opus-4-6

Table 2: Model identifiers and access dates of the APIs used in the experiment.

D The BToM Model and Its Alternatives

D.1 Observer model and retrospective inference

The main text describes the agent model that BToM inverts to solve the food truck task from the perspective of an observer (see Baker et al. for details of the BToM model). The observer model treats the agent's reward function (r) over the three trucks and its belief (b_t) about the occupant of G2 as latent variables to be inferred jointly from the observed trajectory. This process incorporates marginalizing over the uncertainty of the agent's prior beliefs, noisy observations, and goal-directed action selections. As the observer sees the full trajectory to the end (T), recovering a belief at an earlier step (t) is a smoothing, or offline, inference problem. Therefore, the posterior factorizes into a forward and a backward term:

$$P(b_t, r \mid s_{1:T}) \propto \underbrace{P(b_t, r|s_{1:t})}_{\text{forward}} \times \underbrace{P(s_{t+1:T} \mid b_t, r)}_{\text{backward}} \tag{2}$$

Here $s_{1:T}$ denotes the observed sequence of agent positions (given the fixed map and truck configuration), T is the final step of the trajectory, and b_t is the agent's belief about the occupant of G2 at step t; world states in Equation 1 are written s_t for consistency with Baker et al. (2017). The forward term carries the evidence available up to the current step ($s_{1:t}$). The backward term is a likelihood over what the agent did next ($s_{t+1:T}$). It is worth stating what does *not* enter: the true world state that the agent eventually observes is not the evidence about b_1, because the agent did not have access to it at $t = 1$. In the example of Check-GoBack scenario, the BToM observer raises the posterior on $b_1(M)$ because the agent's actions to pass G1 (truck K) to check G2 (truck L) and return to G1 are more probable under an initial belief that truck M was at G2—not because of what was in fact parked at G2. This distinction is important because humans and BToM succeeded in separating the agent's false initial belief that M would be at G2 from the fact that the agent saw L instead at G2.

D.2 Alternative models

Among the four alternative observer models, TrueBelief, NoCost, and MotionHeuristic follow Baker et al., whereas Hindsight is introduced here. The first two remove one component of the BToM structure; MotionHeuristic dispenses with the generative account altogether.

1. TrueBelief solves a Markov Decision Process (MDP) rather than a POMDP. The agent is assumed to observe the world state without uncertainty just as the observer does, so belief plays no role and nothing remains to be inferred about the truck that is never seen (i.e., the always absent truck M).
2. NoCost solves the POMDP but sets the cost of movement to zero. Only the terminal reward differentiates the agent's trajectories, so the intermediate tradeoffs (e.g., length of a detour to G2) carry no information about the agent's desire and belief.
3. MotionHeuristic performs no inference based on a generative model. Observer ratings are predicted from changes in the agent's distance to each truck K, L, M (e.g., whether the

agent moves toward or away from it). To adapt this account to our dynamic evaluation, we introduce a time index t. Let I_t and O_t represent the cumulative average distance changes from the agent to each target (Into) and to the other target (Other) up to step t, respectively. The stepwise predicted rating (V_t) is:

$$V_t = \omega_{into} \cdot I_t + \omega_{other} \cdot O_t + \omega_{intercept} \tag{3}$$

The three weights are estimated by least squares against the human ratings, over the complete trajectory data. This is a simplified version of the motion heuristic of Baker et al. (2017), which used a larger set of weights.

D.3 The Hindsight model

Algorithm 1 Hindsight model algorithm

Input: Trajectory length L, True world state y_{true}, Visual time t_{vis}, Forward log-probs fwd
Parameter: Hindsight bias strength λ
Output: Smoothed log-probs $smooth$

for $\tau = L$ **down to** 1 **do**
  $bwd_\tau(\tau) \leftarrow 0$
  $smooth_\tau(\tau) \leftarrow fwd_\tau + bwd_\tau(\tau)$
  for $t = \tau - 1$ **down to** 1 **do**
    $bwd_\tau(t) \leftarrow$ Compute by agent's transition model
    $penalty \leftarrow \mathbb{I}(\tau \geq t_{vis}) \cdot \lambda \cdot \ln(b_t(y_{true}) + \epsilon)$
    $smooth_\tau(t) \leftarrow fwd_t + bwd_\tau(t) + penalty$
  end for
end for
return $smooth$

Hindsight is a suboptimal model introduced to describe the unique belief dynamics that the LLMs showed in the Every-step condition. It leaves the smoothing distribution (Equation 2) intact except at one point. It evaluates a past belief at step t from a later vantage point τ (the step the Every-step query has reached), iterating backward over all $t < \tau$ (Algorithm 1). However, once τ has passed the step at which the agent visually confirms which truck is at G2 (t_{vis}), the model applies a negative penalty to any past belief hypothesis (b_t) that assigned a near-zero probability to this true world state (y_{true}). The penalty is scaled by a Hindsight bias parameter λ, which dictates how aggressively the post-observation knowledge overrides the agent's prior belief state. This mechanism formalizes the pattern of LLMs that emerges after a hidden world state is revealed. Note that the penalty depends on the true world state y_{true}, which the agent had not observed at step t; this is precisely the information that the normative smoothing in Equation 2 excludes, whereas the agent's later actions enter both models through the backward term.

D.4 Parameter fitting

Model	Optimized β
BToM	2.5
TrueBelief	9.0
NoCost	2.5
MotionHeuristic	N/A
Hindsight	3.5

Table 3: Parameter values for each model fitted to the human ratings. For Hindsight, β was estimated with λ fixed at 2.0; MotionHeuristic has no β (see text).

The softmax parameter β, which scales how deterministically the agent follows the optimal policy, was fitted to the human ratings for BToM and its ablated models using the bootstrap cross-validation procedure of Baker et al. (Table 3). Note that MotionHeuristic has no such parameter; its three regression weights are estimated as described above (Equation 3).

In case of Hindsight penalty parameter λ, the penalty vanishes when $\lambda = 0$ and the Hindsight model becomes identical to BToM. λ is therefore not a parameter that improves the fit to human behavioral data; it is the axis along which the stepwise LLM responses in the Every-step condition are different from the human benchmark, and we fixed it at $\lambda = 2.0$ for the analyses with $\beta = 3.5$. Because λ was fixed rather than fitted and Hindsight was formulated after inspecting the Every-step data, the Hindsight results should be read as descriptive; fitting λ per LLM with cross-validation is an important next step.

SUPPLEMENTARY RESULTS

E CORRELATION ANALYSIS

Table 4: End-step correlations of each LLM with human ratings (Pearson r pooled over 73 scenarios $\times$ 3 options; 95% intervals from 2,000 scenario-level bootstrap resamples). Group: stratification used for the diagnostic analyses (Section 4.1). Every-step: whether the model returned valid stepwise output. The human–BToM row gives the bootstrap cross-validated values of Baker et al. (2017).

Model	Group	r_{desire} [95% CI]	r_{belief} [95% CI]	Every-step
GPT-4o	Worse	0.09 [−0.10, 0.27]	−0.01 [−0.18, 0.14]	excluded
GPT-5.4	Worse	0.15 [−0.03, 0.32]	0.23 [−0.00, 0.42]	valid
GPT-o4-mini	Worse	0.35 [0.20, 0.50]	0.76 [0.69, 0.82]	valid
Gemini-2.5-Flash	Worse	0.62 [0.50, 0.72]	0.33 [0.16, 0.48]	valid
DeepSeek-V3.2	Worse	0.35 [0.19, 0.51]	−0.28 [−0.40, −0.16]	excluded
Claude-4.6-Sonnet	Worse	0.73 [0.65, 0.81]	0.69 [0.60, 0.76]	valid
Gemini-2.5-Pro	Better	0.78 [0.70, 0.83]	0.90 [0.86, 0.93]	valid
DeepSeek-R1	Better	0.81 [0.75, 0.86]	0.59 [0.49, 0.69]	valid
DeepSeek-V4-Pro	Better	0.84 [0.79, 0.88]	0.73 [0.64, 0.80]	valid
Claude-4.6-Opus	Better	0.86 [0.81, 0.91]	0.76 [0.61, 0.86]	valid
Human–BToM	–	0.91	0.78 [0.70, 0.85]	n/a

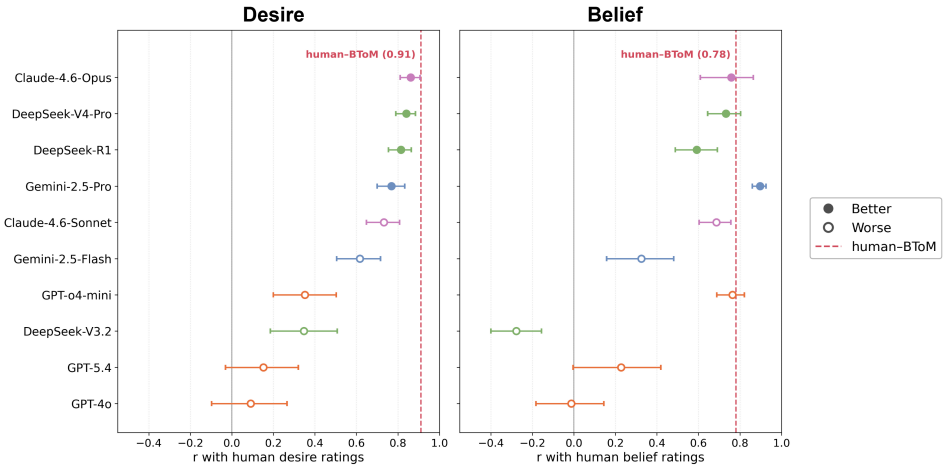


Figure 6: Alignment between LLM and human ratings in the End-step condition. Each point is the correlation between ratings of LLMs and humans, pooled over scenarios and rating options; error bars are 2.5–97.5 percentile intervals from 2,000 bootstrap resamples of scenarios, with the three food-truck options of a scenario resampled together. LLMs were categorized by whether their correlations to human ratings are higher or lower than the correlations to all other alternative models such as NoCost, TrueBelief, MotionHeuristic, and Hindsight (Filled markers for "Better" and open markers for "Worse"). Models are ordered by their desire correlation, and the same order is used in belief. The dashed line marks the human–BToM correlation.

E.1 HUMAN-LLM CORRELATIONS

To summarize how closely each LLM matched the human ratings in the End-step condition, Figure 6 plots the correlations with the human-BToM alignment for a comparison ($r_{desire} = 0.91$, $r_{belief} = 0.78$). Models in the Better group (filled markers) occupy the upper region of both panels. For example, Gemini-2.5-Pro is the one case where the belief interval lies above human-BToM baseline ([0.86, 0.93] against [0.70, 0.85]). For the Worse group (open markers) at the other extreme, the intervals of GPT-4o and GPT-5.4 include zero in both dimensions, while DeepSeek-V3.2's belief interval lies entirely below zero.

However, the partition is not clear-cut: Gemini-2.5-Pro (Better; $r_{desire} = 0.78$) and Claude-4.6-Sonnet (Worse; $r_{desire} = 0.73$) are adjacent in desire, and in belief GPT-o4-mini (Worse; $r_{belief} = 0.76$) even exceeds DeepSeek-R1 (Better; $r_{belief} = 0.59$). Overall, no single ordering of the ten models is consistent across both panels, and the rank orders of the desire and belief correlations agree only moderately ($\rho_{\text{Spearman}} = 0.65$), suggesting that different computational processes underlie desire and belief inferences. Therefore, for each LLM, we asked which of the available models or data best reproduces that LLM's ratings—the human data, or one of the four alternative observer models. We labeled an LLM "Better" when the human data was the best match in this comparison in at least one of the two dimensions, and "Worse" otherwise. Note that the LLM categorization was intended only for convenience of the following diagnostic analyses (e.g., Every-step condition), not a final verdict for LLMs' ToM capacity.

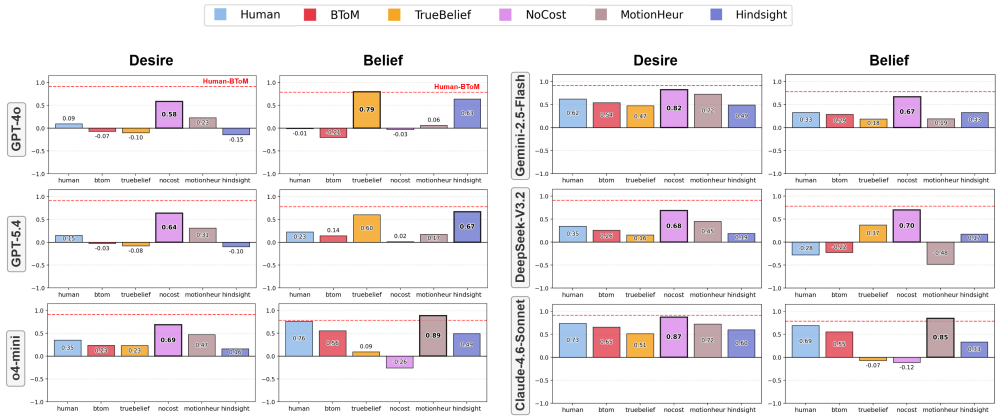


Figure 7: Correlations between each worse-performing LLM and the reference models in the End-step condition. Rows are the six Worse LLMs; the left and right columns show desire and belief ratings. The computational model that best accounts for each LLM's rating (i.e., the tallest bar in each panel) is drawn with a bold outline. These panels decompose the group-level aggregate result shown in Figure 4a of the main text.

E.2 LLM-ABLATION CORRELATIONS

The ablation analysis on End-step responses of Worse LLMs showed that they were consistently well aligned with NoCost for desire. But for belief, different alternative models explained best for different LLMs. As the main text reported only the aggregate summary across the Worse LLMs, Figure 7 shows results from the same analysis for individual LLMs.

F VARIANCE PARTITIONING ANALYSIS

In the main text, we reported only the summary values for the better-performing models, so Figure 8 provides the full variance partitioning result in the End-step condition. Two results are worth noting: first, in desire, the unique variance attributable to the BToM model varied even within the Better group (0.034 to 0.174) and is substantially higher than in the Worse models, though NoCost still accounts for the largest variance in two of the four. Second, in belief, no single account

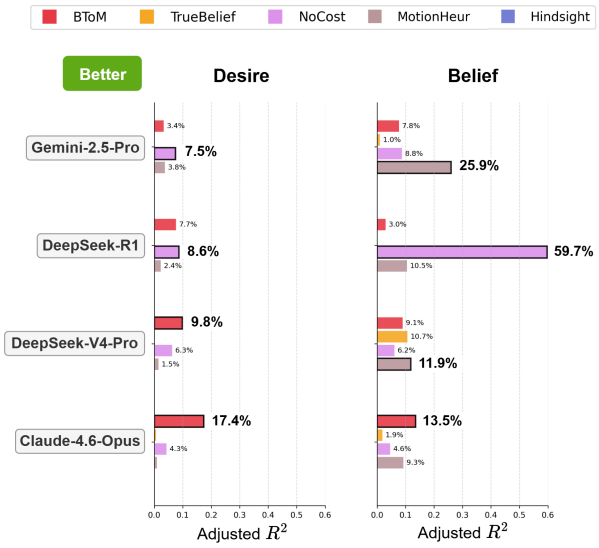


Figure 8: Variance partitioning for the better-performing LLMs in the End-step condition. Rows are the four Better LLMs; the left and right columns show desire and belief ratings. The best explanatory model for each LLM is shown with a bold outline.

dominates the explanatory power: MotionHeuristic explains most of the unique variance for Gemini-2.5-Pro and DeepSeek-V4-Pro, NoCost does so for DeepSeek-R1, and only Claude-4.6-Opus is best explained by BToM in both desire and belief. This heterogeneity of belief inference is also consistent with the differences we observed from Worse models in Figure 7. This is in stark contrast to a relatively homogeneous model comparison result for desire inferences, suggesting again that belief inference poses a different computational problem to LLMs compared to desire.

G Reasoning Logs

We additionally collected reasoning traces from Claude-4.6-Opus with extended reasoning enabled (five runs on the Check-GoBack scenarios) to test whether the belief failure in Every-step condition stems from insufficient deliberation. Three tendencies recur in the excerpts of Table 5.

1. The model repeatedly infers about the absent truck M, revising its estimate—sometimes settling near the normative answer and sometimes not.
2. It also goes beyond what the prompt asks for: alongside the stepwise ratings it tracks a confidence that grows with accumulated evidence, and reads the agent's pause as the agent deliberating between options.
3. The Hindsight bias that the Every-step variance partitioning analysis pointed out is observed in an explicit form: the model inferred that as L was observed at G2 at $t = 4$, the agent at $t = 1$ would believe L to be there with certainty.

These are single traces from one model under a non-default setting, and the Claude 4 API returns summarized rather than full reasoning, so we treat them as qualitative examples consistent with the quantitative results rather than as independent evidence for them.

H Every-step Prompting and Chain-of-Thought

Chain-of-Thought (CoT) prompting and our Every-step design both leverage test-time computation to elicit step-by-step deductions, but they differ in their representations. CoT relies on fluid natural language, which has a risk of introducing semantic ambiguity. In contrast, Every-step condition enforces explicit subgoals of outputting rating scores and therefore helps to reduce the linguistic vagueness. Furthermore, the Hindsight bias in the Every-step condition is opposite to "unfaithful

Type	Every-step reasoning log
Loop for truck M	I'm estimating K around 5-6, L around 3-4, and M potentially similar to K or higher based on their willingness to investigate. (...) Now I'm reconsidering M's preference score. (...) I'll bump M up slightly to account for this exploratory behavior, though **it's somewhat speculative without more direct evidence.** (...) Now I'm reconsidering whether M should be rated higher. (...) Still, without direct evidence, **I'll keep M in the 4-5 range** and stick with my current assessment.
Calculating Certainty	Now I need to think about how the evidence accumulates over time. Early on at step 1, there's barely any signal: they just started moving right, which could mean anything. By steps 5-6, it's becoming clearer (...) So my confidence in **these preference values should grow as the evidence mounts**, but the underlying scores themselves stay constant. (...) The pause at step 15 looks like they're **deliberating** between the two options they've now seen.
Belief in Hindsight	At t=4, the student observes L at Spot 2, which confirms that **L was there all along since trucks don't move during the day**. So working backward, at t=1 the student would now **believe L was at Spot 2 with certainty** (...) initially at t=1 itself, before seeing anything, the student would have had **roughly equal priors** across the three possibilities.

Table 5: Reasoning log examples from the Every-step condition for Claude-4.6-Opus (extended reasoning enabled) on the Check-GoBack scenarios. Bold marks are for the author's emphasis and ellipses mark omitted intervening text within the same trace.

conclusions" typically observed in CoT where final responses contradict preceding valid reasoning (Lu et al., 2025). Future studies are warranted to compare CoT and the Every-step prompting in detail and to understand their commonalities and differences.