OBToM: Ordinal Bayesian Theory of Mind from Likelihood Judgments of Large Language Models

Anonymous authors
Paper under double-blind review

Abstract

Large language models (LLMs) often answer theory-of-mind questions inconsistently when asked to reason directly about mental states. Methods that embed an LLM within Bayesian inverse planning instead require verbalized probabilities, which are poorly calibrated, or token log-probabilities, which many deployed models do not expose. We propose Ordinal Bayesian Theory of Mind (OBToM), which separates likelihood judgment from probability estimation. A story is parsed into discrete variables on a fixed mental-state graph linking prior belief and perception to belief, and belief and goal to action. The LLM is asked only ordinal questions: given two hypothetical conditions and outcomes, it judges whether one conditional likelihood is somewhat or much larger, whether the two are approximately equal, or whether the story provides no basis for judgment. An ordered-logit model converts these judgments into conditional probability tables with a shared low-rank parameterization, so a few fixed comparisons per story also constrain table rows that were never queried. Belief revision is gated by perceptual access: an agent who missed a change keeps the earlier belief. Mental states are then inferred through exact Bayesian inference over the graph, using a maximum a posteriori estimate of the tables or Hamiltonian Monte Carlo samples. We compare OBToM with direct prompting, perspective-taking, timeline, thought-tracing, and verbalized- and token-probability inverse-planning methods on the same backbone. With either estimator, OBToM is the most accurate on BigToM, FANToM, ExploreToM, and first- and second-order ToMi, and significantly so against every baseline on BigToM, ToMi, and ExploreToM. It uses at most five LLM requests per question, against 24.5 to 177.3 for thought-tracing and automated inverse planning.

1 Introduction

Theory of mind is the ability to attribute beliefs, goals, and percepts to other agents and to predict their behavior from these attributions (Premack & Woodruff, 1978). Its canonical test is the false-belief task, in which an agent who is absent while an object is moved should later search where the object used to be; second-order versions ask what one agent thinks another agent believes (Le et al., 2019; Wu et al., 2023). Large language models (LLMs) pass many such tests, yet they fail on small variations, such as making a container transparent (Ullman, 2023; Shapira et al., 2024), and on stories generated to require tracking many actions and nested beliefs (Sclar et al., 2025). Their answers to questions about the same story can also be inconsistent: a model can state a character's belief correctly and then mispredict the behavior that follows from it (Gu et al., 2026), or answer the same belief question differently in different formats (Kim et al., 2023). Under direct prompting, a single generation must track who perceived which event, revise beliefs, and map beliefs to actions, with no constraint that these steps form a consistent account of the agent's mind.

Bayesian theory of mind provides such an account (Baker et al., 2009; 2017). It treats beliefs and goals as latent causes of action and perception as the channel through which beliefs change. Every question about a story is answered by inverting the same generative model. Recent systems embed an LLM in this computation and obtain the likelihood of an action or utterance under each hypothesis from token probabilities (Jin et al., 2024; Shi et al., 2025; Zhang et al., 2025b) or from verbalized numbers and likelihood categories (Gelpí et al., 2025; Kim et al., 2025). Token probabilities are not

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.

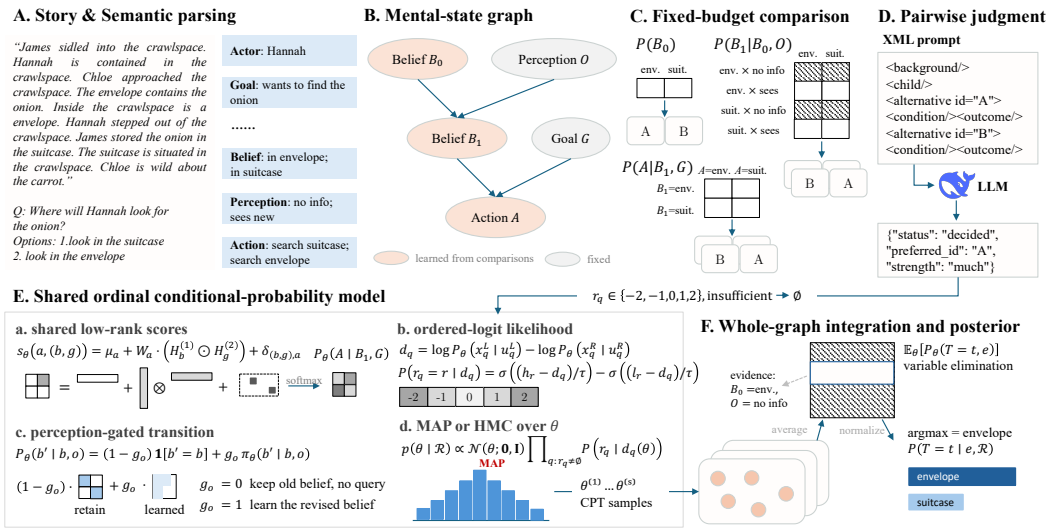


Figure 1: OBToM on a ToMi story: parsing (A), the fixed graph with learned (shaded) and fixed factors (B), the comparison budget, where hatched rows for unperceived changes are never queried (C), an XML comparison and its JSON decision (D), the ordinal CPT model with its MAP and HMC estimates (E), and exact inference over the graph (F).

exposed by many deployed models and lose calibration after post-training (OpenAI, 2023; Tian et al., 2023), and verbalized probabilities are overconfident, cluster on a few round values, and violate the probability axioms (Xiong et al., 2024; Zhu & Griffiths, 2024). Inference then multiplies numbers whose absolute scale the LLM does not report reliably.

Bayesian inference over a discrete graph needs conditional probability tables (CPTs), and a CPT can be constrained without eliciting any number. Deciding which of two cases is more likely is easier and more consistent than rating each case on an absolute scale, for people (Christiano et al., 2017) and for LLMs used as judges (Liusie et al., 2024; Liu et al., 2024). A comparison such as "if Hannah believes the onion is in the envelope, is she more likely to look in the envelope or in the suitcase?" draws on commonsense knowledge of how beliefs guide action, and answering it requires no probability scale.

We propose Ordinal Bayesian Theory of Mind (OBToM), in which the LLM judges relative likelihoods and an explicit statistical model computes probabilities (Fig. 1). A story is parsed into a fixed mental-state graph, and the LLM compares hypothetical conditions paired with outcomes, using five comparisons for a story with one change. An ordered-logit model converts the graded answers into a posterior over CPTs with a shared low-rank parameterization. Exact Bayesian inference over the graph then answers each question using either the maximum a posteriori (MAP) tables or posterior samples from Hamiltonian Monte Carlo (HMC). Two design choices tie the answers to one account of the agent's mind. The LLM never sees the question, and perception enters only as evidence, so true-belief and false-belief versions of a story draw on the same judgments. Belief revision is gated by perceptual access, so an agent who missed a change keeps the previous belief exactly.

We formulate LLM-assisted Bayesian theory of mind using ordinal judgments of conditional likelihood, including abstention, and an observation model that links these judgments to probabilities (Sections 3 and 4.2). We develop a shared low-rank ordinal CPT model with normalized cross-context comparisons and perception-gated transitions, whose posterior enters exact graph inference as a MAP estimate or HMC samples (Sections 4.3 to 4.5). We evaluate OBToM on BigToM, first- and second-order ParaphrasedToMi, FANToM, ExploreToM, and MMToM-QA, with DeepSeek-V4-Flash as every method's backbone. With either estimator, OBToM obtains the highest accuracy on BigToM, FANToM, ExploreToM, and first- and second-order ToMi, with significant gains over all baselines on BigToM, ToMi, and ExploreToM. For example, accuracy reaches 93.6% (MAP) and 92.6% (HMC) on BigToM against 84.7% for the best baseline, and 78.2% and 79.7% against 66.2% on second-order ToMi.

2 Related Work

Theory of mind in large language models. Text benchmarks test false-belief reasoning about moved objects (ToMi; Le et al., 2019) and paraphrases of these stories (Sclar et al., 2023), causal templates that separate belief and action questions (BigToM; Gandhi et al., 2023), conversations whose participants leave and rejoin (FANToM; Kim et al., 2023), and adversarially generated stories with many actions and nested beliefs (ExploreToM; Sclar et al., 2025). Evaluations differ in their assessment of LLMs' theory-of-mind abilities (Kosinski, 2024; Strachan et al., 2024). Prompting methods restructure the reasoning: SimToM filters the story to what the agent knows (Wilf et al., 2024), TimeToM orders each character's beliefs on a timeline (Hou et al., 2024), and SymbolicToM and DEL-ToM track beliefs in explicit graphs or in epistemic logic (Sclar et al., 2023; Wu et al., 2025). Their answers remain LLM-generated text, and no probabilistic model ties answers to different questions together.

Bayesian inverse planning with large language models. Bayesian theory of mind explains mental-state attribution as inverse planning in a generative model of rational action and perception (Baker et al., 2009; 2017). Goodman et al. (2006) expressed false-belief reasoning as a causal Bayesian network in which belief follows the world only under perceptual access; OBToM keeps this structure and learns from LLM judgments what an agent concludes when it has access. Recent work combines Bayesian theory of mind with LLMs in two ways (Table 3 in Appendix A). LLMs either translate language into planning domains or epistemic formulas for a symbolic inverse planner (Ying et al., 2025b;a), or supply inverse-planning likelihoods. These likelihoods come from token probabilities of LLMs, including fine-tuned models, in BIP-ALM, LIMP, CLIPS, AutoToM, and a scalable Bayesian planner (Jin et al., 2024; Shi et al., 2025; Zhi-Xuan et al., 2024; Zhang et al., 2025b;a), and from verbalized probabilities or likelihood categories in LAIP and ThoughtTracing (Gelpí et al., 2025; Kim et al., 2025). GRAIL avoids numeric outputs by asking whether a player's belief should rise or fall and mapping the answer to a fixed prior (Rahimirad et al., 2026). OBToM asks for graded comparisons of conditional likelihoods, fits a posterior over the tables with an explicit noise model, and answers by exact inference over the graph.

Probabilities from large language models and from comparisons. Token probabilities are often miscalibrated (Jiang et al., 2021), more so after post-training (OpenAI, 2023; Tian et al., 2023). Probabilities stated in text need not match those the model assigns (Hu & Levy, 2023), verbalized confidence does not reflect the model's intrinsic uncertainty (Xiong et al., 2024), and answers shift with prompt format and option position (Sclar et al., 2024; Zheng et al., 2024; Wang et al., 2024). Paired comparison is more stable: Thurstone and Bradley-Terry models, extended to ties and graded preferences, recover latent scales from paired judgments (Thurstone, 1927; Bradley & Terry, 1952; Rao & Kupper, 1967; Agresti, 1992), and pairwise prompting improves LLM evaluation, ranking, and preference elicitation (Liusie et al., 2024; Qin et al., 2024; Mazeika et al., 2025). Probabilistic models derived from LLMs mostly elicit numbers, including the CPTs of Bayesian networks (Requeima et al., 2024; Capstick et al., 2025; Nafar et al., 2026), and BIRD fits conditional probabilities to verbal labels mapped to fixed values (Feng et al., 2025). Elicitation of Bayesian networks from human experts met the same difficulty (Druzdzel & van der Gaag, 2000) and used ratio-scale pairwise comparisons (Monti & Carenini, 2000) and order statements imposed as hard constraints on CPTs (Feelders & van der Gaag, 2006; Niculescu et al., 2006). OBToM treats each comparison as a noisy graded observation under an ordered-logit likelihood (McCullagh, 1980), extends it to comparisons across conditions, and couples the rows of a table through a low-rank logistic parameterization (Rijmen, 2008; Kolda & Bader, 2009).

3 Problem Formulation

Mental-state graphs. An instance consists of a story, a question Q about an agent, and a finite set of candidate answers. We represent the mental states needed to answer Q with a directed acyclic graph $\mathcal{G}$ over discrete variables $X_1, \ldots, X_n$, where X_i takes values in a finite set $\mathcal{X}_i$ of natural-language descriptions, such as "Hannah believes the onion is in the envelope." The joint distribution is $P(x_1, \ldots, x_n) = \prod_i P(x_i \mid x_{\mathrm{pa}(i)})$, and each factor is a conditional probability table (CPT) with one row per parent configuration $u \in \mathcal{U}_i = \prod_{j \in \mathrm{pa}(i)} \mathcal{X}_j$. The edges come from a fixed template (Fig. 1B). For a story with one relevant change, the agent's belief before the change B_0 and perception

of the change O determine the belief after the change B_1, which together with the goal G determines the next action A:

$$P(b_0, o, b_1, g, a) = P(b_0)\,P(o)\,P(b_1 \mid b_0, o)\,P(g)\,P(a \mid b_1, g). \tag{1}$$

Belief states are candidate locations or propositions, perception states describe what the agent perceived of the change, and action states are candidate actions. A story with several relevant events chains the transition, $B_0 \to B_1 \to B_2 \to \cdots$, with one perception variable O_k for the k-th event. A second-order question such as "where does X think that Y will search?" keeps the same factors with re-mapped roles: B_k is X's model of Y's belief, and O_k records whether X observed Y witness event k. Stories share the template and differ only in their states and texts.

Queries and factors. A parser maps the story and Q to the graph, a target variable T whose states are the candidate answers, and evidence e, an assignment to observed variables. Forward questions observe the perception, and possibly the initial belief, and ask for the current belief or the next action; backward questions observe an action and ask for the belief or goal behind it. The parser never reads the answer key; we implement it with rules for benchmarks generated from templates and, for conversations, with one LLM call that extracts who is present at each turn (Appendix B). The perception prior is uniform, the goal is the one stated in the story unless the question asks about it, and transition rows for changes the agent could not perceive retain the previous belief (Section 4.4). The LLM informs $P(B_0)$, the rows of $P(B_1 \mid B_0, O)$ with perceptual access, and $P(A \mid B_1, G)$.

Ordinal likelihood judgments. The LLM is accessed only through comparison queries. For a learned factor with child X and parents U, a query q presents two alternatives (u_q^L, x_q^L) and (u_q^R, x_q^R), each a parent configuration with a child outcome, and asks which of $P(X = x_q^L \mid U = u_q^L)$ and $P(X = x_q^R \mid U = u_q^R)$ is larger. A within-row query ($u_q^L = u_q^R$) compares two outcomes, and a cross-context query ($x_q^L = x_q^R$) compares one outcome under two conditions. The response indicates that one alternative is somewhat or much more likely, that the two are about equal, or that the story provides insufficient information. We encode it as $r_q \in \{-2, -1, 0, 1, 2\} \cup \{\varnothing\}$, where positive values favor the left alternative and $\varnothing$ denotes abstention. The query set $\mathcal{Q}$ depends only on the graph and is fixed before any response is collected; for Eq. (1) with binary states it has five elements.

Objective. Given the graph, $\mathcal{Q}$, and the responses $\mathcal{R} = \{(q, r_q)\}_{q \in \mathcal{Q}}$, the task is to compute the posterior $P(T = t \mid e, \mathcal{R})$ for every state t of T and to answer with its mode. Unlike inference with known parameters, this task uses responses that constrain probabilities only through noisy inequalities of uncertain strength. Some responses are abstentions, and coverage is sparse: a row receives at most a few judgments, a cross-context response relates two rows only through their normalized likelihoods, and some rows and outcome pairs receive no judgment.

4 Method

Algorithm 1 summarizes the procedure for one story. Its LLM calls depend only on the graph, so a single fit answers every question that shares the graph.

4.1 Eliciting Ordinal Judgments

Prompts. Each query is an XML prompt (Fig. 1D) with a background, the child variable, and two alternatives A and B, each listing its parent states and its outcome. The prior-belief factor sees the initial situation; the revision factor also sees the change, stated as an event in the world that the agent learns about only through the assigned perception; the action factor sees the agent's task and circumstances but not the change, which can affect the action only through the current belief. Assigned parent states override conflicting implications of the story, so the LLM can evaluate conditions that differ from reality, such as a false belief. The question is never shown, and the evidence that separates true-belief and false-belief versions of a story enters only through e.

Responses. The LLM returns JSON with a status (decided, equal, or insufficient information), a preferred alternative, and a strength (somewhat or much), validated against a strict schema. Equality requires a positive reason to regard the likelihoods as approximately equal, whereas insufficient

Algorithm 1 OBToM for one story and the questions that share its graph.

Input: a story with questions $Q_1, \ldots, Q_M$; an LLM; estimator (MAP or HMC); number of HMC draws S

$(\mathcal{G}, \{g_o\}, \{(T_m, e_m)\}_{m=1}^{M}) \leftarrow \textsc{Parse}(\text{story}, Q_1, \ldots, Q_M)$

$\mathcal{Q} \leftarrow \textsc{Queries}(\mathcal{G}, \{g_o\})$ *fixed in advance; rows with $g_o = 0$ skipped*

$\mathcal{R} \leftarrow \{(q, r_q)\}_{q \in \mathcal{Q}}$ with $r_q = \textsc{Encode}\big(\text{LLM}(\textsc{Prompt}(q))\big)$ *cached; A/B order alternates*

for each learned factor i of $\mathcal{G}$, with responses $\mathcal{R}_i \subseteq \mathcal{R}$ **do**

  $\ell_i(\theta_i) \leftarrow \log \mathcal{N}(\theta_i; \mathbf{0}, \boldsymbol{I}) + \sum_{(q, r_q) \in \mathcal{R}_i,\, r_q \neq \varnothing} \log P\big(r_q \mid d_q(\theta_i)\big)$ *Eqs. (2) to (4)*

  $\hat{\theta}_i \leftarrow$ best of eight L-BFGS-B maximizations of ℓ_i *MAP estimate*

  $\Theta_i \leftarrow \{\hat{\theta}_i\}$ for MAP, or S draws of HMC(ℓ_i) started near $\hat{\theta}_i$ for HMC

for $m = 1, \ldots, M$ **do** *$\theta^{(s)}$ collects the s-th element of every Θ_i*

  $\alpha_t \leftarrow \sum_s P_{\theta^{(s)}}(T_m = t, e_m)$ for every t, by variable elimination

  output $P(T_m = t \mid e_m, \mathcal{R}) = \alpha_t / \sum_{t'} \alpha_{t'}$ *Eq. (5)*

information applies when the story permits no comparison, a distinction also used in human preference elicitation (Christiano et al., 2017). Decisions map to $r_q \in \{\pm 1, \pm 2\}$ in canonical orientation, equality to 0, and abstention to $\varnothing$. LLMs favor particular answer positions and labels regardless of content, including in pairwise judgments (Zheng et al., 2023; 2024; Wang et al., 2024), so the display order alternates across the queries of a factor.

Query design. Queries are enumerated from the graph before any response is collected. By default, a learned row with K outcomes receives all $K(K-1)/2$ outcome pairs, and rows fixed by the perception gate are skipped. For the action factor, an anchored cross-context design asks about all outcome pairs under one anchor belief and compares each of the first $K-1$ outcomes between the anchor and every other belief. This links the rows with $K(K-1)/2 + (N-1)(K-1)$ queries for N beliefs, no more than the $NK(K-1)/2$ within-row pairs. A story with one change and binary states thus receives five queries. Identical prompts are asked once, so a prompt for a factor shared by several graphs of a story costs one call.

4.2 Ordered-Logit Observation Model

For query q, let $d_q(\theta) = \log P_\theta(x_q^L \mid u_q^L) - \log P_\theta(x_q^R \mid u_q^R)$, where θ collects the parameters of the factor. A response is an ordered-logit observation of d_q (McCullagh, 1980), as in models of graded paired comparisons with ties (Rao & Kupper, 1967; Agresti, 1992),

$$P(r_q = r \mid d_q) = \sigma\big((h_r - d_q)/\tau\big) - \sigma\big((l_r - d_q)/\tau\big), \qquad r \in \{-2, -1, 0, 1, 2\}, \tag{2}$$

where σ is the logistic function, $\tau = 1$, and (l_r, h_r) are adjacent cutpoints from $(-\infty, -1.5, -0.35, 0.35, 1.5, \infty)$. "Much more likely" then corresponds to a likelihood ratio above $\exp(1.5) \approx 4.5$, "somewhat more likely" a ratio between 1.4 and 4.5, and "equal" a ratio within a factor of 1.4, each boundary blurred by logistic noise; abstentions add no term. The cutpoints and the prior of Section 4.3 set the scale that ordinal judgments lack, and both are fixed across benchmarks. In a within-row query the row normalizer cancels and d_q is a difference of scores, as in paired-comparison models (Thurstone, 1927; Bradley & Terry, 1952); a cross-context query keeps both normalizers, since raw scores of different rows are not comparable (Appendix F). Responses are conditionally independent given θ, and each distinct prompt contributes one term.

4.3 Shared Low-Rank Conditional Probability Tables

Let a learned factor have child X with K states and parents $U_1, \ldots, U_J$. OBToM scores outcome x in parent configuration $u = (u_1, \ldots, u_J)$ as

$$s_\theta(x, u) = \mu_x + \sum_{c=1}^{D} W_{xc} \prod_{j=1}^{J} H_{u_j c}^{(j)} + \delta_{u,x}, \qquad P_\theta(x \mid u) = \text{softmax}_x\big(s_\theta(\cdot, u)\big), \tag{3}$$

where μ is a child bias, the second term is a rank-D CP decomposition of the interaction between the child and its parents (Kolda & Bader, 2009), and δ is a local residual. The embedding $H_{u_j}^{(j)}$ of a parent state enters every row that contains this state, so a judgment about one row updates parameters shared by all rows containing that parent state, including rows and outcome pairs that were not queried. The residual lets a row deviate from the shared structure when its judgments require it. Parameters are

whitened, with a standard normal prior and fixed scales (Table 5), and $D = 1$; a factor without parents has one row of free logits. The independent-row ablation $s_\theta(x, u) = \eta_{u,x}$ uses the same judgments and inference procedure, but has no parameters shared across rows.

4.4 Perception-Gated Belief Revision

A belief should change only through what the agent perceives (Goodman et al., 2006), and OBToM builds this principle into the belief transition:

$$P_\theta(b' \mid b, o) = (1 - g_o)\,\mathbf{1}[b' = b] + g_o\,\pi_\theta(b' \mid b, o), \tag{4}$$

where $g_o \in [0, 1]$ is the perception gate of state o, the indicator $\mathbf{1}[b' = b]$ retains the previous belief, and $\pi_\theta(b' \mid b, o) \propto \exp\big(s_\theta(b', (b, o)) + \kappa\,\mathbf{1}[b' = b]\big)$ is a learned revision with a persistence offset $\kappa \geq 0$ toward the previous belief b. A row with $g_o = 0$ keeps the previous belief exactly and is never queried: an agent who missed the change still holds the earlier belief, whatever the LLM knows about the world. A row with $g_o = 1$ is learned from the LLM's judgments about what the perceived event implies. Gates come from the parsed perceptual access, such as who was in the room during a move or present during a turn, never from answer options. The likelihood and the predictions use the same mixture, with gradients through the responsibility of the learned component (Appendix C). A false-belief answer thus follows from the prior belief, exact retention, and the action factor, with no revision judgment involved.

For a second-order question the gate encodes witnessed access: X credits Y with a revision only when X observed Y witness the event, and otherwise X attributes the old belief to Y, even if X knows about the change. In stories with many events, a witnessed event, such as a message about another object, often has no bearing on the queried belief. For these chains we set $\kappa = 2$, so a revision row without informative judgments keeps the previous belief with log-odds 2 (probability 0.88 for binary beliefs), which a "much more likely" judgment can overturn. Stories with one change use $\kappa = 0$.

4.5 Parameter Estimation and Answering

The posterior $p(\theta \mid \mathcal{R}) \propto \mathcal{N}(\theta; \mathbf{0}, \boldsymbol{I}) \prod_{q:\, r_q \neq \varnothing} P(r_q \mid d_q(\theta))$ combines the prior with the likelihood of Eq. (2) and factorizes over factors. OBToM answers questions by exact inference over the graph under one of two estimates of the tables. The MAP estimate $\hat{\theta}$ maximizes the posterior; we compute it with L-BFGS-B (Byrd et al., 1995) from eight starting points, seven of them random, because the CP product has a stationary point at the origin. The HMC estimate integrates over the posterior using draws $\theta^{(1)}, \ldots, \theta^{(S)}$ from HMC (Neal, 2011) with analytic gradients and dual-averaging step-size adaptation (Hoffman & Gelman, 2014). The chains start near $\hat{\theta}$ (800 draws per factor from two chains; Appendix C). MAP uses the single most probable set of tables without accounting for parameter uncertainty, whereas HMC averages over tables in proportion to their posterior probability, at the cost of 1,600 sampling iterations per factor. Both estimates answer with

$$P(T = t \mid e, \mathcal{R}) = \frac{\mathbb{E}_{\theta \sim p(\theta \mid \mathcal{R})}\big[P_\theta(T = t, e)\big]}{\mathbb{E}_{\theta \sim p(\theta \mid \mathcal{R})}\big[P_\theta(e)\big]} \approx \frac{\sum_{s=1}^{S} P_{\theta^{(s)}}(T = t, e)}{\sum_{s=1}^{S} P_{\theta^{(s)}}(e)}, \tag{5}$$

where MAP uses $S = 1$ and $\theta^{(1)} = \hat{\theta}$, and the joint probabilities are computed exactly by variable elimination (Koller & Friedman, 2009) for all draws in one pass. For HMC, averaging the joint before conditioning weights each draw by the probability $P_\theta(e)$ that it assigns to the evidence, as Bayes' rule requires once e is observed; averaging the per-draw conditionals $P_\theta(T = t \mid e)$ would drop this weight (Appendix F). The answer is the mode of Eq. (5).

5 Experiments

5.1 Setup

Benchmarks. We evaluate on six settings (Tables 1 and 2): BigToM (Gandhi et al., 2023); the first-order and second-order questions of the ParaphrasedToMi validation split (Sclar et al., 2023), denoted ToMi; FANToM with short conversations and multiple-choice belief questions (Kim et al., 2023); ExploreToM (Sclar et al., 2025); and the text version of MMToM-QA (Jin et al., 2024).

Table 1: Accuracy (%) on held-out questions; n is the number of questions. ToMi (1st) denotes the first-order questions of ParaphrasedToMi; second-order results are in Table 2. On ExploreToM, ThoughtTracing and AutoToM were run on 383 and 500 of the held-out questions. The best result in each column is in bold; ° marks baselines whose difference from both OBToM rows is not significant after Holm correction (Appendix E).

Method	BigToM	ToMi (1st)	FANToM	ExploreToM	MMToM-QA
n	2,160	1,902	1,428	9,035	500
Direct	74.26	75.60	52.45	57.06	40.80
SimToM	72.92	84.17	69.68°	57.39	55.60°
TimeToM	71.25	67.67	69.47°	57.53	44.40
ThoughtTracing	72.08	75.34	53.92	54.83	**62.80**°
LAIP	84.72	80.44	71.15°	73.18	45.40
AutoToM	75.69	80.65	59.59	59.40	57.20°
OBToM-MAP	**93.56**	**95.06**	**71.85**	**82.40**	54.80
OBToM-HMC	92.55	94.90	71.57	81.56	56.80

The first five test false beliefs, including second-order beliefs in ToMi, FANToM, and ExploreToM. MMToM-QA asks instead which belief or goal explains a person's actions in an apartment, and OBToM compiles it into the graph $B_1, G \rightarrow A$ with the observed action sequence as evidence (Appendix B). A small development split of each benchmark was used to design the prompts, and we report accuracy on the remaining held-out questions.

Baselines and backbone. We compare with direct prompting, SimToM, TimeToM, ThoughtTracing, LAIP, and AutoToM (Section 2). We run the official code of ThoughtTracing and AutoToM and the official prompts of SimToM, and we reimplement TimeToM and LAIP from their papers. Every method uses DeepSeek-V4-Flash as its backbone. For OBToM we report the shared model with both estimators of Section 4.5, MAP and HMC, using the settings of Table 5 on every benchmark.

5.2 Main Results

Table 1 reports accuracy on BigToM, first-order ToMi, FANToM, ExploreToM, and MMToM-QA. With either estimator, OBToM is the most accurate method on the four false-belief benchmarks in the table. Across the two estimators, its margin over the strongest baseline is 7.8 to 8.8 points on BigToM, 10.7 to 10.9 on first-order ToMi, and 8.4 to 9.2 on ExploreToM; on FANToM, LAIP comes within 0.4 to 0.7 points. We test each difference with a paired bootstrap over stories and apply Holm correction to the 36 comparisons of each estimator (Appendix E); MAP and HMC then lead to the same conclusions. OBToM is significantly more accurate than every baseline on BigToM, first-order ToMi, and ExploreToM. On FANToM, its gains over direct prompting, ThoughtTracing, and AutoToM are significant, while the differences from LAIP, SimToM, and TimeToM are not. The strongest baseline varies by benchmark: LAIP leads on three, and SimToM on first-order ToMi. Direct prompting stays below 60% on FANToM, ExploreToM, and MMToM-QA.

On MMToM-QA, ThoughtTracing is the most accurate method with 62.8%, followed by AutoToM with 57.2%, while OBToM reaches 56.8% with HMC and 54.8% with MAP. OBToM is significantly more accurate than direct prompting, TimeToM, and LAIP on MMToM-QA, and its differences from SimToM, ThoughtTracing, and AutoToM are not significant after correction; ThoughtTracing's lead of 8.0 points over MAP has an uncorrected p of 0.014. Across the six settings of Tables 1 and 2, the two estimators differ by at most 2.0 points, and neither dominates: MAP is more accurate on four settings and HMC on the other two.

5.3 Second-Order Beliefs

Table 2 isolates the second-order questions of ToMi. OBToM answers them with the first-order factors re-mapped to X's model of Y's belief and gated by whether X saw Y witness the move (Section 4.4), using the same prompt templates as for first-order questions. It reaches 78.2% with MAP and 79.7% with HMC, 12.0 to 13.5 points above LAIP, the strongest baseline, and 14.1 to 15.6 points above AutoToM. All differences from the baselines are significant after correction. Every method loses

Table 2: Second-order questions of ParaphrasedToMi ("where does X think that Y will search?", $n = 1{,}902$), with first-order accuracy from Table 1 and the change from first to second order. The two question sets come from different stories of the same split. The best second-order result is in bold.

Method	ToMi (1st)	ToMi (2nd)	Change
Direct	75.60	54.84	−20.76
SimToM	84.17	59.25	−24.92
TimeToM	67.67	53.52	−14.15
ThoughtTracing	75.34	59.31	−16.03
LAIP	80.44	66.19	−14.25
AutoToM	80.65	64.09	−16.56
OBToM-MAP	95.06	78.23	−16.83
OBToM-HMC	94.90	**79.71**	−15.19

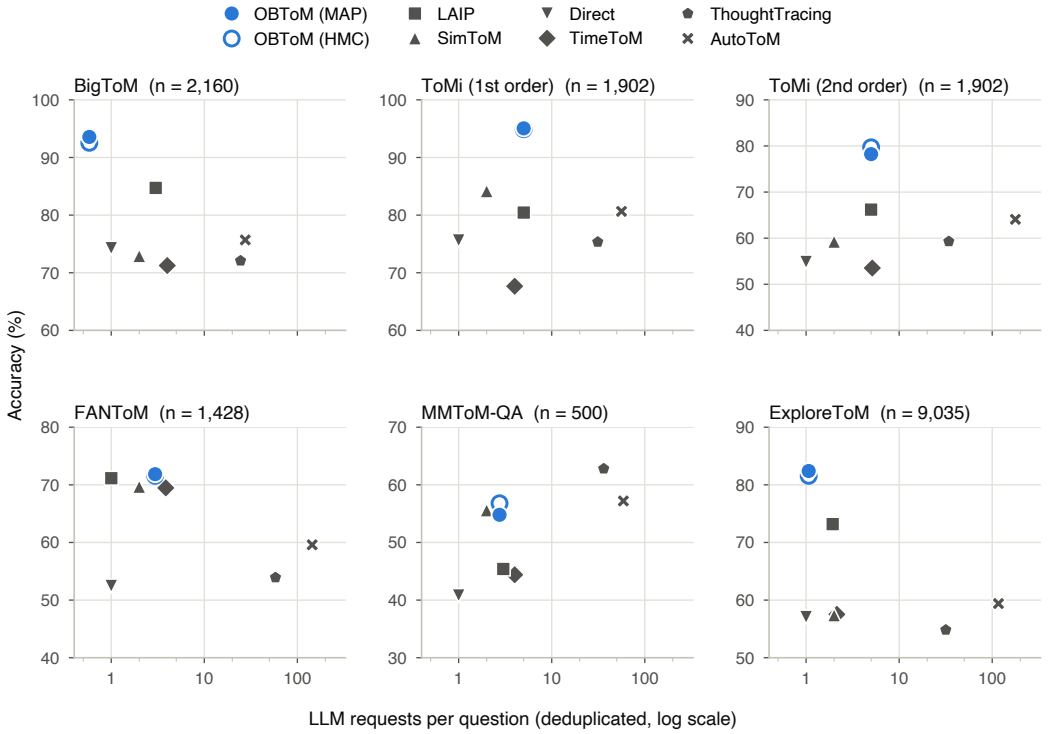


Figure 2: Accuracy on held-out questions against the average number of LLM requests per question (log scale; identical requests within a question counted once). Points toward the upper left are better. Blue marks OBToM, filled for MAP and open for HMC; both use the same judgments and therefore the same number of requests. Gray marks the baselines.

accuracy from first-order to second-order questions, by 14.2 to 24.9 points. SimToM, which filters the story to the perspective of one agent, loses the most and falls from first to fourth among the baselines. OBToM loses 15.2 points with HMC and 16.8 with MAP, within the range of the baselines, and its lead over the strongest baseline grows from 10.7 to 10.9 points on first-order questions to 12.0 to 13.5 points on second-order questions.

5.4 Number of LLM Requests

Figure 2 plots accuracy against the average number of LLM requests per question, and Table 6 in Appendix D lists the numbers. The requests of OBToM are fixed by the graph and shared by all questions about it, so BigToM, where six questions share each graph, needs fewer than one request

per question. Across the six settings, OBToM uses 0.58 to 5.00 requests per question, compared with 1.00 to 5.14 for direct prompting, SimToM, TimeToM, and LAIP. ThoughtTracing and AutoToM make repeated calls to generate and score hypotheses, using 24.5 to 58.2 and 27.5 to 177.3 requests per question, respectively. On the five false-belief settings, no method uses fewer requests and achieves higher accuracy than OBToM; on BigToM, it uses the fewest requests and achieves the highest accuracy. On second-order ToMi, AutoToM uses 177.3 requests per question and reaches 64.1%, whereas OBToM uses 5.0 and reaches 78.2% to 79.7%. ThoughtTracing is the most accurate method only on MMToM-QA, where it uses about 13 times as many requests as OBToM (36.2 against 2.75).

6 Conclusion

We presented OBToM, which separates the two roles that LLM-based theory of mind usually assigns to one model. The LLM supplies ordinal comparisons of conditional likelihoods, from which an ordered-logit model estimates a posterior over conditional probability tables with shared low-rank rows and perception-gated belief revision. Exact Bayesian inference over a fixed mental-state graph then answers each question using MAP tables or HMC samples. With DeepSeek-V4-Flash as the backbone of every method, OBToM was the most accurate method on BigToM, FANToM, ExploreToM, and first- and second-order ToMi with either estimator, significantly so against every baseline on BigToM, ToMi, and ExploreToM. Its lead over the strongest baseline grew from first-order to second-order questions, and it used 0.58 to 5.00 LLM requests per question, against 24.5 to 177.3 for the hypothesis-based methods.

The method needs neither token probabilities nor numeric reports from the LLM, so it applies to deployed models that expose neither. Its answers to different questions about a story are posteriors of one joint distribution, so the true-belief and false-belief versions of a story draw on the same judgments, and a false belief follows from exact retention without any judgment about it. Each answer can also be traced to a graph, a short list of judgments, and a posterior. Integrating parameter uncertainty with HMC did not consistently improve accuracy over MAP, which estimates the tables by optimization.

The parsers are written for benchmarks generated from templates, and for conversations they use one LLM call to extract who hears each turn; open-ended narratives will require a general semantic parser, whose errors the graph would inherit. The template covers the beliefs, goals, perceptions, and actions of one agent and their second-order re-mapping. Emotions, intentions, deception, or deeper nesting would need new factors, while exact inference restricts the graphs to a small number of joint states. The cutpoints that translate verbal strength into log-likelihood ratios are fixed across benchmarks, and judgments are treated as conditionally independent given the tables. OBToM judges a long action sequence as a single outcome. All methods were evaluated with a single backbone. Learning the cutpoints for each model, splitting long action sequences into steps, and pairing the ordinal interface with a learned semantic parser are direct extensions. OBToM also returns a probability for every candidate answer, and HMC carries the uncertainty of the tables into these probabilities; evaluating their calibration is left to future work.

AI Use Statement

We used generative AI tools for four tasks: cleaning and formatting benchmark data, formulating some of the mathematical claims and their proofs, polishing the text of the paper, and searching the literature. The authors reviewed all AI-assisted output. The LLM judgments used by OBToM and by the baselines are part of the method and the experiments and are described in Sections 4 and 5. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

Reproducibility Statement

We provide the complete source code and experimental run logs in the supplementary material. Section 4 and Appendix C specify the observation model, parameterization, priors, cutpoints, and inference procedures, and Table 5 lists the hyperparameters. All settings are shared across benchmarks

except the persistence offset in Table 4. Appendix B describes how each benchmark is compiled into graphs, queries, and evidence.

REFERENCES

Alan Agresti. Analysis of ordinal paired comparison data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 41(2):287–297, 1992.

Chris L. Baker, Rebecca Saxe, and Joshua B. Tenenbaum. Action understanding as inverse planning. *Cognition*, 113(3):329–349, 2009.

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.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Richard H. Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

Alexander Capstick, Rahul Krishnan, and Payam Barnaghi. AutoElicit: Using large language models for expert prior elicitation in predictive modelling. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 6746–6777, 2025.

Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4299–4307, 2017.

Marek J. Druzdzel and Linda C. van der Gaag. Building probabilistic networks: "Where do the numbers come from?" Guest editors' introduction. *IEEE Transactions on Knowledge and Data Engineering*, 12(4):481–486, 2000.

Ad Feelders and Linda C. van der Gaag. Learning Bayesian network parameters under order constraints. *International Journal of Approximate Reasoning*, 42(1–2):37–53, 2006.

Yu Feng, Ben Zhou, Weidong Lin, and Dan Roth. BIRD: A trustworthy Bayesian inference framework for large language models. In *International Conference on Learning Representations*, pp. 8961–8989, 2025.

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*, volume 36, pp. 13518–13529, 2023.

Andrew Gelman and Donald B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992.

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.

Noah D. Goodman, Chris L. Baker, Elizabeth Baraff Bonawitz, Vikash K. Mansinghka, Alison Gopnik, Henry Wellman, Laura Schulz, and Joshua B. Tenenbaum. Intuitive theories of mind: A rational approach to false belief. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 28, pp. 1382–1387, 2006.

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. In *International Conference on Learning Representations*, pp. 50252–50297, 2026.

Matthew D. Hoffman and Andrew Gelman. The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(47):1593–1623, 2014.

Guiyang Hou, Wenqi Zhang, Yongliang Shen, Linjuan Wu, and Weiming Lu. TimeToM: Temporal space is the key to unlocking the door of large language models' theory-of-mind. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 11532–11547, 2024.

Jennifer Hu and Roger Levy. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5040–5060, 2023.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. How can we know when language models know? On the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977, 2021.

Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer D. Ullman, Antonio Torralba, Joshua B. 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.

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. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14397–14413, 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. In *Second Conference on Language Modeling (COLM)*, 2025.

Tamara G. Kolda and Brett W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3): 455–500, 2009.

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

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

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5872–5877, 2019.

Yinhong Liu, Han Zhou, Zhijiang Guo, Ehsan Shareghi, Ivan Vulić, Anna Korhonen, and Nigel Collier. Aligning with human judgement: The role of pairwise preference in large language model evaluators. In *Conference on Language Modeling*, 2024.

Adian Liusie, Potsawee Manakul, and Mark Gales. LLM comparative assessment: Zero-shot NLG evaluation through pairwise comparisons using large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 139–151, 2024.

Mantas Mazeika, Xuwang Yin, Rishub Tamirisa, Jaehyuk Lim, Bruce W. Lee, Richard Ren, Long Phan, Norman Mu, Oliver Zhang, and Dan Hendrycks. Utility engineering: Analyzing and controlling emergent value systems in AIs. In *Advances in Neural Information Processing Systems*, volume 38, pp. 140822–140868, 2025.

Peter McCullagh. Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(2):109–127, 1980.

Stefano Monti and Giuseppe Carenini. Dealing with the expert inconsistency in probability elicitation. *IEEE Transactions on Knowledge and Data Engineering*, 12(4):499–508, 2000.

Aliakbar Nafar, K. Brent Venable, Zijun Cui, and Parisa Kordjamshidi. Extracting probabilistic knowledge from large language models for Bayesian network parameterization. *Transactions on Machine Learning Research*, 2026.

Radford M. Neal. MCMC using Hamiltonian dynamics. In Steve Brooks, Andrew Gelman, Galin L. Jones, and Xiao-Li Meng (eds.), *Handbook of Markov Chain Monte Carlo*, chapter 5, pp. 113–162. Chapman and Hall/CRC, 2011.

Radu Stefan Niculescu, Tom M. Mitchell, and R. Bharat Rao. Bayesian network learning with parameter constraints. *Journal of Machine Learning Research*, 7(50):1357–1383, 2006.

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, and Michael Bendersky. Large language models are effective text rankers with pairwise ranking prompting. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 1504–1518, 2024.

Shahab Rahimirad, Guven Gergerli, Lucia Romero, Angela Qian, Matthew Lyle Olson, Simon Stepputtis, and Joseph Campbell. Bayesian social deduction with graph-informed language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 32415–32440, 2026.

P. V. Rao and L. L. Kupper. Ties in paired-comparison experiments: A generalization of the Bradley-Terry model. *Journal of the American Statistical Association*, 62(317):194–204, 1967.

James Requeima, John Bronskill, Dami Choi, Richard E. Turner, and David Duvenaud. LLM processes: Numerical predictive distributions conditioned on natural language. In *Advances in Neural Information Processing Systems*, volume 37, pp. 109609–109671, 2024.

Frank Rijmen. Bayesian networks with a logistic regression model for the conditional probabilities. *International Journal of Approximate Reasoning*, 48(2):659–666, 2008.

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 (Volume 1: Long Papers)*, pp. 13960–13980, 2023.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *International Conference on Learning Representations*, pp. 25055–25083, 2024.

Melanie Sclar, Jane Dwivedi-Yu, Maryam Fazel-Zarandi, Yulia Tsvetkov, Yonatan Bisk, Yejin Choi, and Asli Celikyilmaz. Explore Theory of Mind: program-guided adversarial data generation for theory of mind reasoning. In *International Conference on Learning Representations*, pp. 67635–67660, 2025.

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 (Volume 1: Long Papers)*, pp. 2257–2273, 2024.

Haojun Shi, Suyu Ye, Xinyu Fang, Chuanyang Jin, Leyla Isik, Yen-Ling Kuo, and Tianmin Shu. MuMA-ToM: Multi-modal multi-agent theory of mind. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 1510–1519, 2025.

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, 2024.

L. L. Thurstone. A law of comparative judgment. *Psychological Review*, 34(4):273–286, 1927.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5433–5442, 2023.

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

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450, 2024.

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.

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

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. 11383–11397, 2025.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *International Conference on Learning Representations*, pp. 23650–23678, 2024.

Lance Ying, Ryan Truong, Katherine M. Collins, Cedegao E. Zhang, Megan Wei, Tyler Brooke-Wilson, Tan Zhi-Xuan, Lionel Wong, and Joshua B. Tenenbaum. Language-informed synthesis of rational agent models for grounded theory-of-mind reasoning on-the-fly. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 12217–12235, 2025a.

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, 2025b.

Chunhui Zhang, Zhongyu Ouyang, Kwonjoon Lee, Nakul Agarwal, Sean Dae Houlihan, Soroush Vosoughi, and Shao-Yuan Lo. Overcoming multi-step complexity in multimodal theory-of-mind reasoning: A scalable Bayesian planner. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 75878–75900, 2025a.

Zhining Zhang, Chuanyang Jin, Mung Yao Jia, Shunchi Zhang, and Tianmin Shu. AutoToM: Scaling model-based mental inference via automated agent modeling. In *Advances in Neural Information Processing Systems*, volume 38, pp. 129659–129699, 2025b.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *International Conference on Learning Representations*, pp. 19426–19454, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623, 2023. Datasets and Benchmarks Track.

Tan Zhi-Xuan, Lance Ying, Vikash Mansinghka, and Joshua B. Tenenbaum. Pragmatic instruction following and goal assistance via cooperative language-guided inverse planning. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pp. 2094–2103, 2024.

Jian-Qiao Zhu and Thomas L. Griffiths. Incoherent probability judgments in large language models. In *Proceedings of the 46th Annual Conference of the Cognitive Science Society*, pp. 906–913, 2024.

Table 3: LLM-based theory-of-mind and Bayesian methods, grouped by the source of their probabilities.

Method	Model of the mind	LLM output entering the computation
SimToM (Wilf et al., 2024)	none; perspective filtering	filtered story, then the answer as text
TimeToM (Hou et al., 2024)	temporal belief chains per character	timeline and belief chains, then the answer as text
SymbolicToM (Sclar et al., 2023)	belief graphs per character	answer from sentences retrieved from the graph
BIP-ALM (Jin et al., 2024)	inverse planning over symbolic states	action likelihoods from token probabilities of a fine-tuned LLM
LIMP (Shi et al., 2025)	inverse multi-agent planning	action and utterance likelihoods from token probabilities
AutoToM (Zhang et al., 2025b)	automatically constructed agent model	each local conditional from the token probability of "likely"
ThoughtTracing (Kim et al., 2025)	weighted free-text hypotheses	one of six verbal likelihood categories mapped to numbers
LAIP (Gelpí et al., 2025)	inverse planning	verbalized prior and likelihood values
BIRD (Feng et al., 2025)	Bayesian model over factors (not ToM)	verbal likelihood labels mapped to fixed values, then fitted
GRAIL (Rahimirad et al., 2026)	factor graph over hidden roles	whether a belief should rise, fall, or stay, mapped to a fixed prior
OBToM (ours)	fixed mental-state graph with perception-gated transitions	graded comparisons of two conditional likelihoods, fitted by an ordered-logit likelihood (MAP or HMC estimate of the tables)

A How Related Methods Obtain Probabilities

Table 3 summarizes how the LLM-based methods discussed in Section 2 represent an agent's mind and which LLM outputs enter the computation. Prompting methods read the answer from generated text. Methods based on inverse planning combine LLM outputs with Bayes' rule, and they differ in how the LLM output is turned into a number.

B Compiling Benchmarks into Mental-State Graphs

Each benchmark has a parser that produces the graph, the candidate states, the perception gates, the target and evidence of each question, and the fixed query set. Apart from the presence extraction for FANToM, the parsers are deterministic programs over the public text of the benchmark and make no LLM call. Candidate states are the answer options of the benchmark or the containers and rooms named in the story; their order is randomized per story, and state names such as `belief_0` carry no information about correctness. Answer keys are read only by the scorer. Table 4 summarizes the compiled inputs.

BigToM. Each BigToM story instantiates a causal template of five sentences (Gandhi et al., 2023): the agent and setting, the agent's goal, the initial situation, the agent's initial belief, and a change in the world, followed by a sentence stating whether the agent perceives the change. We use sentences one to three as the setup, sentence two as the goal, and sentence five as the change. The template's two perception sentences, in which the agent does or does not perceive the change, are the states of O with gates 1 and 0, and the two belief options of the template are the states of B_0 and B_1. Each template yields two graphs. In one, the action states are the two behaviors described by the template, which serve as evidence for backward-belief questions; in the other, they are the options of forward-action questions. The two graphs share their prior and revision prompts, which are asked once. Forward-belief and forward-action questions observe O and ask for B_1 or A. Backward-belief questions observe A, marginalize O under its uniform prior, and ask for B_1. Every question type appears with and without the initial-belief sentence; when the sentence is shown, B_0 is observed, and

Table 4: Compiled inputs. ToMi denotes the ParaphrasedToMi validation split, and FANToM the short conversations with multiple-choice belief questions. A query is one comparison request. Identical prompts are cached, so the number of distinct prompts equals the number of comparison calls to the LLM. FANToM also makes one presence-extraction call per conversation (369 in total), which Table 6 includes. κ is the persistence offset of Eq. (4).

Benchmark	Graph	Questions	Queries	Distinct prompts	Prompts / question	κ
BigToM	one change	2,400	2,000	1,395	0.58	0
ToMi, first order	one change	2,002	10,010	10,010	5.00	0
ToMi, second order	re-mapped	2,002	10,010	10,010	5.00	0
FANToM	dialogue chain	1,528	4,156	4,156	2.72	2
ExploreToM	event chain	9,584	52,564	10,222	1.07	2
MMToM-QA	belief and goal	600	1,650	1,650	2.75	n/a

otherwise it is marginalized under the elicited prior. The action factor uses the anchored cross-context design of Section 4.1.

ParaphrasedToMi. ParaphrasedToMi (Sclar et al., 2023) renders ToMi stories (Le et al., 2019) with a finite set of paraphrase templates for each type of event. The parser matches every sentence against these templates and recovers entries, exits, object placements, and the single move of the queried object. The belief and action states are the initial container of the object and its destination. The setup contains the sentences before the move and, when it follows the move, the sentence that locates the destination container; the move is the change, and later sentences determine only the evidence. For a first-order question about agent X, the perception states are "X sees that the object is now in the destination" (gate 1) and "X does not see the move and is not in the room afterwards" (gate 0). X perceives the move if X is in the room when it happens, performs it, or enters the room afterwards. The initial placement is public, so B_0 is observed. A second-order question, "where does X think that Y searches for the object?", uses the re-mapped graph of Section 3 with three perception states: X sees Y learn the new location (gate 1), X sees the new location without evidence about Y (gate 0), and X observes nothing (gate 0). X sees Y learn the new location if both are present at the move, if Y re-enters the room while X is there, or if X re-enters the room and finds Y there. The action factor uses within-row pairs, which cost the same as the anchored design for binary states.

FANToM. FANToM (Kim et al., 2023) asks about the beliefs of participants who leave and rejoin a conversation. We use the short conversations and the multiple-choice belief questions, whose two options become the belief states verbatim. The parser needs to know who could hear each turn. For each distinct conversation, a single LLM call, which sees the turns but no question or answer, returns when each participant arrives and leaves; the output is checked against the turns, and every speaker must be present while speaking. For 12 of the 1,540 questions the extraction fails this check on all three attempts; these questions are not compiled, and every method is evaluated on the compiled questions. A first-order question about X uses X's presence as the mask, and a second-order question about X's belief concerning Y uses the conjunction of both presences. The turns are grouped into maximal runs with a constant mask value. The turns heard in the first run form the background of the prior factor, which otherwise states explicitly that nothing has been heard. Each later run is an event: an absent run is exact retention, and a present run is a learned revision whose background lists the turns heard so far and whose change lists the new turns.

ExploreToM. ExploreToM (Sclar et al., 2025) generates stories from a domain-specific language of actions: entering and leaving rooms, moving objects between containers or rooms, public and private communication, changes of object states, and modifiers for secret witnessing and distraction. We parse the released story structures with the generator's sentence templates and replay the actions to record, for each action, which agents witnessed it attentively or in secret and which agents could see each other witnessing it. For a question about a proposition, namely the container or room of an object, knowledge of a topic, or the state of an object, the relevant events are the actions that bear on it: moves, messages on the topic, state changes, and entries into a room where the object lies in view. The graph is a chain with one transition per relevant event. A first-order chain for X gates an event on X witnessing it, and a second-order chain for X's belief about Y gates it on X witnessing the

Table 5: Hyperparameters, shared by all benchmarks except κ (Table 4).

Component	Setting
Ordered logit	cutpoints $(-1.5, -0.35, 0.35, 1.5)$, noise scale $\tau = 1$
Shared tables	rank $D = 1$; prior scales 0.25 (μ), 1.4 (W, H), 0.2 (δ)
Root factors, independent rows	free logits, prior scale 1.5
Belief transition	exact retention when $g_o = 0$; gates $g_o \in \{0, 1\}$ from the parse
MAP	L-BFGS-B from 8 starts (zero and seven draws from $\mathcal{N}(\mathbf{0}, 0.7^2\boldsymbol{I})$), at most 800 iterations
HMC	2 chains with 400 warmup and 400 retained draws each; initial step size 0.08, adapted by dual averaging to acceptance 0.95 and jittered by $\pm 10\%$; trajectory length uniform on $\{8, \ldots, 16\}$ leapfrog steps; chains start at the best MAP solutions plus $\mathcal{N}(\mathbf{0}, 0.15^2\boldsymbol{I})$ noise
Seed	7 for every graph

event while Y visibly witnesses it. Belief states are the containers or rooms involved in the relevant events, or knowing and not knowing (believing and not believing) for knowledge and state questions. Search questions add the action factor with within-row pairs, and knowledge and state questions query the last belief. Backgrounds refer to the belief holder as "the witness" and contain only the rooms, the people, and the candidate states, so all chains of a story that pass through the same event share its prompts, which reduces 52,564 queries to 10,222 distinct prompts. We evaluate the 9,584 mental-state questions with at least two candidate answers; the remaining 3,725 questions ask about the state of the world or have a single candidate.

MMToM-QA. The text version of MMToM-QA (Jin et al., 2024) describes the state of an apartment and a person's actions up to a point before a decisive step, and asks a two-option question about the person's belief given a goal, the goal given a belief, or the goal alone. We compile each question into the graph $B_1, G \to A$, whose action variable has two states: the observed sequence and any other course of action. The two options are the states of the queried variable, and a belief or goal stated in the question is a single fixed state. For goal-only questions the belief is latent, with one state for accurate and one for inaccurate knowledge of what the containers hold. The priors over beliefs and goals are uniform. The queries compare the observed sequence with its alternative under one anchor configuration and compare the likelihood of the observed sequence between the anchor and every other configuration of belief and goal. The background is the state description verbatim, and the answer is the posterior of the queried variable given the observed sequence.

C Model and Inference Details

Gradients. Table 5 lists all settings. The objective is the negative log posterior of Section 4.5, and all gradients are analytic. Writing $z_\ell = (l_r - d)/\tau$ and $z_h = (h_r - d)/\tau$, the ordered-logit term has

$$\log P(r \mid d) = \log\big(\sigma(z_h) - \sigma(z_\ell)\big), \qquad \frac{\partial}{\partial d} \log P(r \mid d) = \frac{\sigma(z_\ell) + \sigma(z_h) - 1}{\tau}, \tag{6}$$

and it is evaluated in a form that stays finite when both $\sigma(z_\ell)$ and $\sigma(z_h)$ round to 0 or 1. For a normalized table, $\partial \log P_\theta(x \mid u)/\partial s_\theta(x', u) = \mathbf{1}[x' = x] - P_\theta(x' \mid u)$. For a gated row of Eq. (4) with learned component $\pi_\theta(b' \mid b, o) \propto \exp\big(s_\theta(b', (b, o)) + \kappa\, \mathbf{1}[b' = b]\big)$,

$$\frac{\partial \log P_\theta(b' \mid b, o)}{\partial s_\theta(x', (b, o))} = \rho_\theta(b' \mid b, o)\, \big(\mathbf{1}[x' = b'] - \pi_\theta(x' \mid b, o)\big), \qquad \rho_\theta(b' \mid b, o) = \frac{g_o\, \pi_\theta(b' \mid b, o)}{P_\theta(b' \mid b, o)}, \tag{7}$$

where ρ_θ is the responsibility of the learned component for outcome b'. The derivatives with respect to μ, W, H, and δ follow from Eq. (3) by the chain rule, including the prior scales of the whitened parameterization.

Sampling and diagnostics. HMC uses a unit mass matrix and a Metropolis correction after each trajectory; rejected states are kept in the chain. The step size is adapted only during warmup and then frozen at the averaged value. We record the acceptance rate, the number of divergent transitions

Table 6: Average number of LLM requests per question; identical requests within a question are counted once. MAP and HMC use the same judgments, and FANToM includes presence extraction (one request per conversation).

Method	BigToM	ToMi (1st)	ToMi (2nd)	FANToM	ExploreToM	MMToM-QA
Direct	1.00	1.00	1.00	1.00	1.00	1.00
SimToM	2.00	2.00	2.00	2.00	2.00	2.00
TimeToM	4.00	4.00	5.14	3.86	2.14	4.00
ThoughtTracing	24.54	31.24	34.19	58.19	31.72	36.20
LAIP	3.01	5.01	5.00	1.00	1.93	3.02
AutoToM	27.54	56.19	177.30	144.16	116.68	58.86
OBToM	0.58	5.00	5.00	2.96	1.07	2.75

(energy error above 1000), and the split $\hat{R}$ (Gelman & Rubin, 1992) over all CPT entries. Diagnostics are computed on CPT entries because the sign and order of CP components can change without changing any table.

Exact inference. All tables carry a leading axis over posterior draws, and fixed tables are broadcast along it. Variables that are neither observed nor queried are summed out by variable elimination with a greedy order that minimizes the size of the resulting factor. Each elimination step is a single tensor contraction over all draws, so Eq. (5) costs one pass per question. The graphs used here are chains or small polytrees, every intermediate factor has at most a few dozen entries per draw, and inference is exact for every draw.

D Number of LLM Requests

Table 6 reports the average numbers of LLM requests plotted in Fig. 2. For OBToM, requests shared by the questions of a graph are counted once and divided by the number of questions.

E Paired Significance Tests

For every benchmark and baseline, we compare OBToM with the baseline on the held-out questions that both answered. A paired bootstrap resamples stories 2,000 times, keeping all questions of a story together, and gives a 95% interval for the difference in accuracy and a two-sided p-value; $p < 0.001$ means that no resample crossed zero. We apply Holm correction to the 36 comparisons of each estimator. Table 7 lists all comparisons. After correction, MAP and HMC lead to the same conclusions: OBToM is significantly more accurate in 30 of the 36 comparisons. The remaining six comparisons, against LAIP, SimToM, and TimeToM on FANToM and SimToM, ThoughtTracing, and AutoToM on MMToM-QA, are not significant.

F Derivations

Averaging order in Eq. (5). Observing e is informative about θ as well as about T. The posterior over parameters given both the judgments and the evidence is $p(\theta \mid \mathcal{R}, e) \propto P_\theta(e)\, p(\theta \mid \mathcal{R})$, because e depends on $\mathcal{R}$ only through θ. Hence

$$P(T = t \mid e, \mathcal{R}) = \int P_\theta(T = t \mid e)\, p(\theta \mid \mathcal{R}, e)\, d\theta = \frac{\int P_\theta(T = t, e)\, p(\theta \mid \mathcal{R})\, d\theta}{\int P_\theta(e)\, p(\theta \mid \mathcal{R})\, d\theta}, \tag{8}$$

which is Eq. (5). The average $\int P_\theta(T = t \mid e)\, p(\theta \mid \mathcal{R})\, d\theta$ agrees with it only when $P_\theta(e)$ does not vary with θ. For two equally weighted draws with $P_{\theta^{(1)}}(e) = 0.9$, $P_{\theta^{(1)}}(T = t \mid e) = 0.9$ and $P_{\theta^{(2)}}(e) = 0.1$, $P_{\theta^{(2)}}(T = t \mid e) = 0.1$, Eq. (5) gives 0.82 whereas the average of conditionals gives 0.5. The difference matters for backward questions, where the observed action is evidence about the table of the action factor.

Table 7: Accuracy of OBToM-MAP and OBToM-HMC minus that of each baseline, in percentage points, with 95% story-level bootstrap intervals and two-sided p-values. Positive values favor OBToM. $^\circ$ marks differences that are not significant after Holm correction over the 36 comparisons of each estimator; all others are significant. n is the number of paired questions.

Benchmark	Baseline	n	MAP	p	HMC	p
BigToM	Direct	2,160	+19.31 [+17.5, +21.2]	<0.001	+18.29 [+16.3, +20.2]	<0.001
	SimToM	2,160	+20.65 [+18.5, +22.8]	<0.001	+19.63 [+17.5, +21.8]	<0.001
	TimeToM	2,160	+22.31 [+20.1, +24.6]	<0.001	+21.30 [+19.2, +23.5]	<0.001
	ThoughtTracing	2,160	+21.48 [+19.4, +23.4]	<0.001	+20.46 [+18.4, +22.5]	<0.001
	LAIP	2,160	+8.84 [+7.2, +10.5]	<0.001	+7.82 [+6.2, +9.5]	<0.001
	AutoToM	2,160	+17.87 [+16.1, +19.9]	<0.001	+16.85 [+14.9, +18.8]	<0.001
ToMi (1st)	Direct	1,902	+19.45 [+17.3, +21.7]	<0.001	+19.30 [+17.1, +21.5]	<0.001
	SimToM	1,902	+10.88 [+8.9, +12.8]	<0.001	+10.73 [+8.8, +12.6]	<0.001
	TimeToM	1,902	+27.39 [+25.0, +29.8]	<0.001	+27.23 [+24.9, +29.7]	<0.001
	ThoughtTracing	1,902	+19.72 [+17.6, +21.9]	<0.001	+19.56 [+17.4, +21.8]	<0.001
	LAIP	1,902	+14.62 [+12.6, +16.7]	<0.001	+14.46 [+12.4, +16.5]	<0.001
	AutoToM	1,902	+14.41 [+12.5, +16.4]	<0.001	+14.25 [+12.3, +16.3]	<0.001
ToMi (2nd)	Direct	1,902	+23.40 [+20.5, +26.2]	<0.001	+24.87 [+21.9, +27.8]	<0.001
	SimToM	1,902	+18.98 [+16.2, +21.9]	<0.001	+20.45 [+17.7, +23.2]	<0.001
	TimeToM	1,902	+24.71 [+21.9, +27.5]	<0.001	+26.18 [+23.4, +29.0]	<0.001
	ThoughtTracing	1,902	+18.93 [+16.0, +21.8]	<0.001	+20.40 [+17.6, +23.3]	<0.001
	LAIP	1,902	+12.04 [+9.2, +14.9]	<0.001	+13.51 [+10.8, +16.2]	<0.001
	AutoToM	1,902	+14.14 [+11.3, +17.0]	<0.001	+15.62 [+12.8, +18.7]	<0.001
FANToM	Direct	1,428	+19.40 [+16.1, +22.6]	<0.001	+19.12 [+15.8, +22.3]	<0.001
	SimToM	1,428	+2.17 [−0.8, +5.3]$^\circ$	0.159	+1.89 [−1.1, +4.9]$^\circ$	0.223
	TimeToM	1,428	+2.38 [−0.8, +5.5]$^\circ$	0.152	+2.10 [−1.1, +5.3]$^\circ$	0.204
	ThoughtTracing	1,428	+17.93 [+14.8, +20.9]	<0.001	+17.65 [+14.6, +20.7]	<0.001
	LAIP	1,428	+0.70 [−2.3, +3.6]$^\circ$	0.703	+0.42 [−2.9, +3.6]$^\circ$	0.820
	AutoToM	1,428	+12.25 [+9.0, +15.8]	<0.001	+11.97 [+8.6, +15.5]	<0.001
ExploreToM	Direct	9,035	+25.35 [+22.4, +28.1]	<0.001	+24.50 [+21.6, +27.3]	<0.001
	SimToM	9,035	+25.01 [+22.5, +27.5]	<0.001	+24.17 [+21.6, +26.7]	<0.001
	TimeToM	9,035	+24.87 [+22.2, +27.5]	<0.001	+24.03 [+21.4, +26.7]	<0.001
	ThoughtTracing	383	+24.80 [+18.9, +30.8]	<0.001	+24.80 [+18.6, +31.0]	<0.001
	LAIP	9,035	+9.22 [+6.7, +11.5]	<0.001	+8.38 [+5.9, +10.7]	<0.001
	AutoToM	500	+20.00 [+13.2, +26.8]	<0.001	+19.00 [+12.4, +25.4]	<0.001
MMToM-QA	Direct	500	+14.00 [+8.0, +19.8]	<0.001	+16.00 [+10.2, +21.6]	<0.001
	SimToM	500	−0.80 [−7.2, +5.4]$^\circ$	0.827	+1.20 [−5.0, +7.4]$^\circ$	0.738
	TimeToM	500	+10.40 [+4.0, +16.6]	0.002	+12.40 [+6.4, +18.4]	<0.001
	ThoughtTracing	500	−8.00 [−14.0, −2.0]$^\circ$	0.014	−6.00 [−12.0, +0.0]$^\circ$	0.053
	LAIP	500	+9.40 [+3.8, +15.2]	0.001	+11.40 [+5.6, +17.4]	<0.001
	AutoToM	500	−2.40 [−8.6, +3.2]$^\circ$	0.488	−0.40 [−6.6, +5.4]$^\circ$	0.984

Normalized cross-context comparisons. A cross-context query compares $P_\theta(x \mid u)$ with $P_\theta(x \mid u')$ for $u \neq u'$. With $Z_\theta(u) = \sum_{x'} \exp s_\theta(x', u)$, the difference d_q of Section 4.2 equals

$$d_q(\theta) = s_\theta(x, u) - s_\theta(x, u') - \log Z_\theta(u) + \log Z_\theta(u'). \tag{9}$$

The score difference alone depends on shifts of each row's scores that leave the table unchanged, so it is not identified by the tables; the normalizers remove this dependence and make d_q a function of the tables alone.

Exact retention and false-belief answers. With $g_o = 0$, $P_\theta(B_1 = b \mid B_0 = b, O = o) = 1$ for every θ. For a forward-action question with evidence $B_0 = b$ and $O = o$, the joint then reduces to $P_\theta(A = a, B_0 = b, O = o) = P_\theta(B_0 = b)\, P(O = o)\, P_\theta(A = a \mid B_1 = b, G = g)$ for the stated goal g, and Eq. (5) becomes a weighted average of the action row for belief b. The answer depends only on the judgments about the action factor and, through the weights $P_\theta(B_0 = b)$, on the prior; no judgment about the revision factor enters it.