Understanding Epistemic Language with a Language-augmented Bayesian Theory of Mind

Lance Ying[1,2*], **Tan Zhi-Xuan**[1*], **Lionel Wong**[1]
Vikash Mansinghka[1], **Joshua B. Tenenbaum**[1]
[1]Massachusetts Institute of Technology, Cambridge, MA
[2]Harvard University, Cambridge, MA

Abstract

How do people understand and evaluate claims about others' beliefs, even though these beliefs cannot be directly observed? In this paper, we introduce a cognitive model of epistemic language interpretation, grounded in Bayesian inferences about other agents' goals, beliefs, and intentions: a language-augmented Bayesian theory-of-mind (LaBToM). By translating natural language into an epistemic "language-of-thought" with grammar-constrained LLM decoding, then evaluating these translations against the inferences produced by inverting a generative model of rational action and perception, LaBToM captures graded plausibility judgments of epistemic claims. We validate our model in an experiment where participants watch an agent navigate a maze to find keys hidden in boxes needed to reach their goal, then rate sentences about the agent's beliefs. In contrast with multimodal LLMs (GPT-4o, Gemini Pro) and ablated models, our model correlates highly with human judgments for a wide range of expressions, including modal language, uncertainty expressions, knowledge claims, likelihood comparisons, and attributions of false belief.

1 Introduction

People regularly use and interpret language about other agents' beliefs, evaluating rich linguistic constructions that may involve claims about what others' consider possible (*"Grace thinks Katie might have eaten the cookie"*), what they find more probable (*"Tom thinks Sam is the most likely to win the election"*), or the relationship of their beliefs to the world (*"John didn't know that today was a holiday"*). But since these beliefs are not directly observable, how do people judge the truth

---

*Equal Contribution.

or acceptability of epistemic claims? Philosophers and linguists have long investigated the semantics of epistemic language (Hintikka, 1962; Partee, 1973; Loar, 1981), offering compositional theories of how epistemic claims relate to the sets of worlds deemed possible by an agent (Von Fintel and Heim, 2011). However, these theories do not explain how people ground epistemic language in its context of utterance. If someone says *"Alice believes it might rain"*, how does her behavior (e.g. bringing an umbrella) render that statement more or less plausible or acceptable?

In this paper, we present a computational model of how people interpret epistemic language in context, grounded in inferences about what other agents believe (Figure 1) — a *Language-augmented Bayesian theory-of-mind* (LaBToM). We build upon *Bayesian theory-of-mind* (BToM), a framework which casts human mentalizing as probabilistic inference over a generative model of rational action and perception (Baker et al., 2017; Jara-Ettinger et al., 2020). We combine this framework with the compositionality afforded by probabilistic extensions of the language-of-thought hypothesis (Piantadosi, 2011; Goodman and Lassiter, 2015), developing an *epistemic language of thought* (ELoT) to represent how others represent the world. We parse natural language to this representation using grammar-constrained sequential Monte Carlo decoding (Loula et al., 2025) of large language models (LLMs), defining a flexible but structured probabilistic mapping from epistemic language to epistemic concepts (Wong et al., 2023). This allows us to quantify the plausibility of epistemic sentences against BToM inferences.

To evaluate our model, we run an experiment where participants watch animations of a player solving a gridworld puzzle called Doors, Keys, & Gems (Zhi-Xuan et al., 2020). In these puzzles, the player has to pick up (single-use) keys that
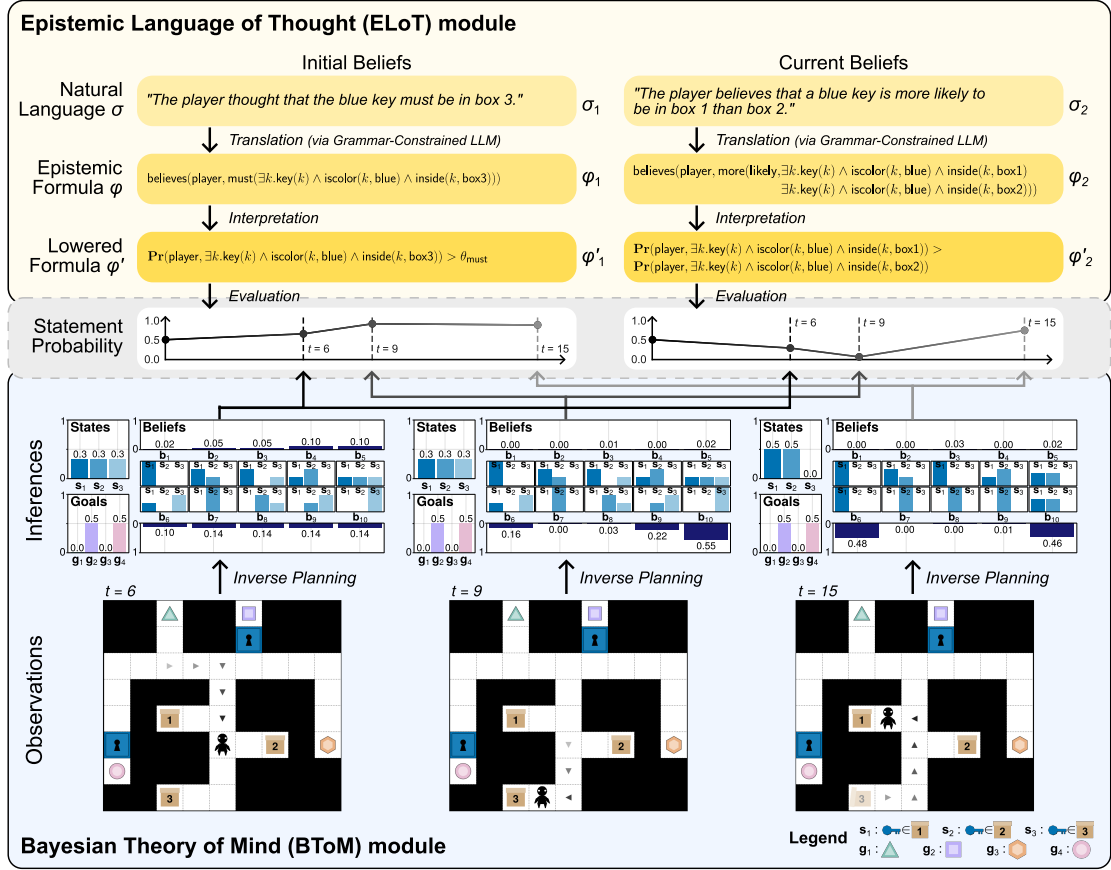
Figure 1: Overview of our model, a Language-augmented Bayesian Theory of Mind (LaBToM). Here, our model evaluates the plausibility of epistemic language about a player character trying to find keys in boxes so as to reach one of four goals (gems with different shapes/colors, $g_1 - g_4$) which may be locked behind doors. **(Top)** We translate natural language statements about the player's initial (σ_1) and current beliefs (σ_2) into an unambiguous *epistemic language of thought (ELoT)* via grammar-constrained LLM parsing. ELoT statements (i.e. epistemic formulas φ) are interpreted with a probability-based semantics (lowered formulas φ'). **(Bottom)** We use our *Bayesian theory-of-mind (BToM)* module to produce inferences (bar charts) about the environment state (top left), the player's goal (bottom left), and the player's belief state (right) at each step t, given observations across time. Each possible belief state b_i is itself a distribution over environment states s_j, where state s_j corresponds to a blue key being in box j. **(Middle)** We evaluate the posterior probability* of each ELoT statement φ from the BToM inferences at each step t (*under a 50-50 prior over statement truth). Statement σ_1 (*"The player thought that the blue key must be in box 3."*) increases in probability from $t=6$ to $t=9$, then stays high, since it becomes clear from the player looking in box 3 at $t=9$ that they *initially* thought a key must be in box 3. Statement σ_2 (*"The player believes that a blue key is more likely to be in box 1 than box 2."*) decreases in probability as the player walks away from box 1 at $t=6$ through $t=9$. However, when the player finds box 3 empty, then moves past box 2 up to box 1 at $t=15$, it becomes much more probable that the player *currently* thinks a key is more likely to be in box 1 rather than box 2.

may be hidden in boxes, using them to unlock doors of the same color to reach one of four valuable gems. The player's beliefs and goals are unknown to our participants, so they must *infer* these mental states. We ask one set of participants to write sentences describing the past and present beliefs of the player, collecting a rich dataset of epistemic language that includes modal verbs, uncertainty expressions, knowledge claims, and descriptions of false beliefs. We task another set of participants with evaluating these statements, asking them to rate how likely a statement is given

the context. We find that the inferences produced by our LaBToM model correlate highly with these human ratings. In contrast, using a lower-level formalization of epistemic claims leads to substantially more parsing errors than our ELoT representation, and ablated BToM models and multimodal LLM baselines fail to explain human sentence judgments. These findings illustrate the importance of a conceptual vocabulary that is aligned with the structure of epistemic language, and the need for a coherent theory-of-mind to systematically interpret such language.

2 Related Work

Model-theoretic formal semantics of epistemic language. Sentences about what agents believe or know have been studied extensively across philosophy and linguistics. Both *epistemic modals* (Yalcin, 2007; Coates, 1987; Egan and Weatherson, 2011; MacFarlane, 2011) and *propositional attitudes* (Frege, 1948; Russell, 1903; Schiffer, 2003) complicate compositional and model-theoretic approaches to assigning truth conditions. Standard treatments associate belief sentences with sets of compatible worlds (Hintikka, 1962; Von Fintel and Heim, 2011), though the role of an agent's mental states in a truth-conditional semantics remains debated (Egan and Weatherson, 2011). More recent work grounds the gradeness of belief claims in probability (Lassiter, 2017; Moss, 2015), which we follow in this paper. Unlike purely model-theoretic approaches, however, our model relates epistemic sentences to a *functional* theory of agents' mental states (Loar, 1981), explaining how people ground these sentences in agent behavior.

Cognitive approaches to (epistemic) language interpretation. Our model draws on accounts of propositional attitudes (A. Fodor, 1981) and word meaning that emphasize the role of mental states (Loar, 1981; Block, 1987; Lake and Murphy, 2023) and the interface between language and conceptual content, as in cognitive or psychosemantics (Fodor, 1987; Lakoff, 1988; Jackendoff, 2003) and functional role semantics (Harman, 1982). We build most closely upon theories that ground linguistic meaning in a (probabilistic) language of thought (Fodor et al., 1975; Goodman and Lassiter, 2015; Wong et al., 2023; Zhang et al., 2023), and models that map sentences to grounded symbolic representations of agents' behavior (Artzi and Zettlemoyer, 2013).

Bayesian Theory-of-Mind. To relate epistemic sentences to (inferred) mental states, we build on the Bayesian Theory-of-Mind framework (Baker et al., 2017; Zhi-Xuan et al., 2022), and related work on epistemic action understanding (Croom et al., 2023; Shvo et al., 2020). BToM provides a functional theory of propositional attitudes like knowledge and belief, leveraging the connection between symbolic representations of the world (used in (inverse) planning (McDermott et al., 1998)) and our representations of others' minds.

Theory-of-Mind in Language Models. With recent advances in LLM capabilities, some researchers have suggested that LLMs might serve as cognitive models (Binz and Schulz, 2023), including as models of theory-of-mind (Strachan et al., 2024), thereby implicitly capturing the semantics of belief sentences (Piantadosi and Hill, 2022). However, while LLMs perform well on some ToM tasks, they do not reliably generalize (Shapira et al., 2024) to richer multi-step (Kim et al., 2023) or multi-modal contexts (Jin et al., 2023; Ying et al., 2024a). Our model instead uses (grammar-constrained) LLMs as a flexible mappings between language and formal meaning representations (Wong et al., 2023; Ying et al., 2023a,b; Zhi-Xuan et al., 2024b), tying language to an *explicit* semantics of epistemic concepts.

3 Computational Model

Our LaBToM model comprises two interlinked modules. The first module (Figure 1, Top), an *epistemic language of thought* (ELoT), models our capacity to compositionally represent the world (including the contents of others' minds) by combining more basic concepts into richer thoughts and expressions (Goodman and Lassiter, 2015), and how we flexibly translate such thoughts from natural language (Wong et al., 2023). The second module (Figure 1, Bottom), a *Bayesian theory-of-mind* (Baker et al., 2017), captures our intuitive inferences about others' minds via Bayesian inference over a generative model of how agents update their beliefs and act towards their goals. Epistemic language understanding can thus be modeled by mapping language into ELoT formulas, which we evaluate against the inferences produced by rational mentalizing.

3.1 Interpreting Belief Sentences with an Epistemic Language of Thought

To represent epistemic concepts in a way that mirrors the structure of natural language, we introduce a formal language (Table 1) as our epistemic language of thought. Drawing upon the formal semantics of epistemic language (Hintikka, 1962; Lassiter, 2010), we adopt a compositional degree-based semantics for epistemic concepts grounded in probability comparisons (Moss, 2015; Lassiter, 2017). This allows us to quantitatively evaluate an epistemic expression using probabilities inferred by our BToM module.

Expression	Type	Arg. Types	Definition
Belief Operators			
$\text{believes}(A, \phi)$	$\mathcal{E}$	$\mathcal{A}, \Phi$	$\mathbf{Pr}(A, \phi) \geq \theta_{\text{believes}}$
$\text{believes}_{\text{modal}}(A, M)$	$\mathcal{E}$	$\mathcal{A}, \mathcal{E}/\mathcal{A}$	$M(A)$
Knowledge Operators			
$\text{knows}_{\text{that}}(A, \phi)$	$\mathcal{E}$	$\mathcal{A}, \Phi$	$\text{believes}(A, \phi) \wedge \phi$
$\text{knows}_{\text{if}}(A, \phi)$	$\mathcal{E}$	$\mathcal{A}, \Phi$	$\text{knows}_{\text{that}}(A, \phi) \vee \text{knows}_{\text{that}}(A, \neg\phi)$
$\text{knows}_{\text{about}}(A, C, \phi)$	$\mathcal{E}$	$\mathcal{A}, \Phi/\mathcal{O}, \Phi/\mathcal{O}$	$\exists x.C(x) \wedge \text{knows}_{\text{that}}(A, \phi(x))$
$\text{not-knows}_{\text{that}}(A, \phi)$	$\mathcal{E}$	$\mathcal{A}, \Phi$	$\neg\text{believes}(A, \phi) \wedge \phi$
Certainty Operators			
$\text{certain}_{\text{that}}(A, \phi)$	$\mathcal{E}$	$\mathcal{A}, \Phi$	$\mathbf{Pr}(A, \phi) \geq \theta_{\text{certain}}$
$\text{certain}_{\text{about}}(A, C, \phi)$	$\mathcal{E}$	$\mathcal{A}, \Phi/\mathcal{O}, \Phi/\mathcal{O}$	$\exists x.C(x) \wedge (\mathbf{Pr}(A, \phi(x)) \geq \theta_{\text{certain}})$
$\text{uncertain}_{\text{if}}(A, \phi, \psi)$	$\mathcal{E}$	$\mathcal{A}, \Phi, \Phi$	$(\mathbf{Pr}(A, \phi) < \theta_{\text{uncertain}}) \wedge (\mathbf{Pr}(A, \psi) < \theta_{\text{uncertain}})$
$\text{uncertain}_{\text{about}}(A, C, \phi)$	$\mathcal{E}$	$\mathcal{A}, \Phi/\mathcal{O}, \Phi/\mathcal{O}$	$\forall x.C(x) \rightarrow (\mathbf{Pr}(A, \phi(x)) < \theta_{\text{uncertain}})$
Modal Verbs			
$\text{could}(\phi)$	$\mathcal{E}/\mathcal{A}$	Φ	$\lambda A.\mathbf{Pr}(A, \phi) \geq \theta_{\text{could}}$
$\text{might}(\phi)$	$\mathcal{E}/\mathcal{A}$	Φ	$\lambda A.\mathbf{Pr}(A, \phi) \geq \theta_{\text{might}}$
$\text{may}(\phi)$	$\mathcal{E}/\mathcal{A}$	Φ	$\lambda A.\mathbf{Pr}(A, \phi) \geq \theta_{\text{may}}$
$\text{should}(\phi)$	$\mathcal{E}/\mathcal{A}$	Φ	$\lambda A.\mathbf{Pr}(A, \phi) \geq \theta_{\text{should}}$
$\text{must}(\phi)$	$\mathcal{E}/\mathcal{A}$	Φ	$\lambda A.\mathbf{Pr}(A, \phi) \geq \theta_{\text{must}}$
Modal Adjectives			
$\text{likely}(\phi)$	$\mathcal{E}/\mathcal{A}$	Φ	$\lambda A.\mathbf{Pr}(A, \phi) \geq \theta_{\text{likely}}$
$\text{degree}(\text{likely}, A, \phi)$	$\Phi_{\mathcal{F}}$	$\mathcal{A}, \Phi$	$\mathbf{Pr}(A, \phi)$
Comparatives			
$\text{more}(P, \phi, \psi)$	$\mathcal{E}/\mathcal{A}$	$\mathcal{P}, \Phi, \Phi$	$\lambda A.\text{degree}(P, A, \phi) > \text{degree}(P, A, \psi)$
$\text{most}_{\text{sup}}(P, O, C, \phi)$	$\mathcal{E}/\mathcal{A}$	$\mathcal{P}, \mathcal{O}, \Phi/\mathcal{O}, \Phi/\mathcal{O}$	$\lambda A.\text{degree}(P, A, \phi(O)) \geq \max_{x:C(x)} \text{degree}(P, A, \phi(x))$
$\text{most}_{\text{str}}(P, \phi)$	$\mathcal{E}/\mathcal{A}$	$\mathcal{P}, \Phi$	$\lambda A.\text{degree}(P, A, \phi) \geq \alpha_{\text{most}} \cdot \theta_P$

(a) Epistemic terms and definitions (unlikely, less and least are omitted due to space limits).

Thresholds θ										Multipliers α
believes	certain	uncertain	likely	unlikely	could	might	may	should	must	most
0.75	0.95	0.70	0.70	0.40	0.20	0.20	0.30	0.80	0.95	1.5

(b) Probability thresholds and multipliers (fitted against human data).

Table 1: Expressions in our epistemic language of thought (ELoT), including **(a)** epistemic terms and **(b)** probability thresholds Θ. ELoT terms may have the following types: $\mathcal{E}$: epistemic formula, Φ: base formula, $\Phi_{\mathcal{F}}$: function term, $\mathcal{P}$: predicate symbol, $\mathcal{A}$: agent, $\mathcal{O}$: object, $\mathcal{X}/\mathcal{Y}$: function from $\mathcal{X} \rightarrow \mathcal{Y}$.

3.1.1 Representing Epistemic Formulas

We first define our non-epistemic base language — a first order logic derived from the Planning Domain Definition Language (McDermott et al., 1998). Our language assumes a set of predicates $\mathcal{P}$ and functions $\mathcal{F}$ used to describe a set of objects $\mathcal{O}$. Predicates can be combined into formulas $\phi \in \Phi$ via logical operators or quantification. For example, *"A key is in box 2"* can be represented as $\exists k.\text{key}(k) \wedge \text{inside}(k, \text{box2})$. Conceptually, a state s of the environment (or an agent's mental representation of state s) is just a large formula: A conjunction of predicates which fully describe the state. We denote the truth value of ϕ in s as $[\![\phi]\!]_s$.

On top of this base language Φ, we introduce epistemic expressions $\varphi \in \mathcal{E}$ to model assertions of belief, knowledge, or modal qualifications (Table 1a). Following Lassiter (2017), the semantics of these expressions are grounded in the probability function $\mathbf{Pr}(A, \phi)$ — the probability assigned by agent A to formula ϕ — and comparisons with term-specific thresholds (Table 1b). For example, the operator $\text{might}(\phi)$ takes a first-order formula $\phi \in \Phi$, returning a function $\lambda A.\mathbf{Pr}(A, \phi) \geq \theta_{\text{might}}$. Combined with the operator $\text{believes}_{\text{modal}}(A, F) = F(A)$, we can express the claim that *"A believes it might be that ϕ"* as $\varphi = \text{believes}_{\text{modal}}(A, \text{might}(\phi))$. This formula φ can be lowered to the probability comparison $\mathbf{Pr}(A, \phi) \geq \theta_{\text{might}}$, which uses θ_{might} as the threshold instead of θ_{believes} due to how our composition rules. We call this latter representation a *lowered formula*.

A key aspect of ELoT expressions is that they represent epistemic concepts at a *similar level of granularity* as our tested language (English), simplifying the mapping to ELoT propositions. In contrast, our lowered representation is just as expressive, but fails to match the structure of natural epistemic language. In our experiments, we show that translating into ELoT significantly improves the accuracy of semantic parsing.

3.1.2 Translating Epistemic Language

Natural language is varied and imprecise, with possibly many sentences σ mapping to the same ELoT formula φ (and vice versa). To handle this diversity, we perform semantic parsing via grammar-constrained LLM decoding (Shin et al., 2021; Scholak et al., 2021; Willard and Louf, 2023), producing a flexible context-sensitive mapping between natural language and ELoT expressions (Wong et al., 2023). Specifically, we use sequential Monte Carlo (SMC)-based grammar-constrained sampling (Loula et al., 2025), since it avoids the failure modes of beam search and greedy token-masking (Lew et al., 2023b). We prompt an LLM (LLaMa 3.1 8B) with example translations $\mathcal{D}$ from English to ELoT and a sentence σ to translate, using SMC to approximate the distribution over completions constrained to the ELoT grammar $G_{\mathcal{E}}$:

$$\varphi \sim P_{\text{LLM}}(\varphi|\sigma, \mathcal{D}, \varphi \in G_{\mathcal{E}}) \qquad (1)$$

We run SMC with n_σ samples. This produces weighted samples $\{(\varphi_i, w_i)\}_{i=1}^{n_\sigma}$, and we use the top-weighted sample as our ELoT translation φ.[1]

3.2 Inferring and Evaluating Beliefs with a Bayesian Theory-of-Mind

Our ELoT semantics reduces belief expressions to the probability $\mathbf{Pr}(A, \phi)$ an agent A assigns to a sentence ϕ. But where does $\mathbf{Pr}(A, \phi)$ come from? This is the function of our BToM module: By modeling the functional role that belief plays in guiding actions, along with the influence of perceptions on beliefs, an observer can infer what the agent thinks based on what the agent sees and does. Following the structure of Partially Observable Markov Decision Processes (POMDPs) (Kaelbling et al., 1998), this theory of approximately rational agency can be formalized as a probabilistic generative model:

$$\textit{Goal Prior:} \quad g \sim P(g) \qquad (2)$$
$$\textit{State Prior:} \quad s_0 \sim P(s_0) \qquad (3)$$
$$\textit{Belief Prior:} \quad b_0 \sim P(b_0|s_0) \qquad (4)$$
$$\textit{State Transition:} \quad s_t \sim P(s_t|s_{t-1}, a_{t-1}) \quad (5)$$
$$\textit{Belief Update:} \quad b_t \sim P(b_t|s_t, b_{t-1}) \qquad (6)$$
$$\textit{Action Selection:} \quad a_t \sim P(a_t|b_t, g) \qquad (7)$$
$$\textit{Observations:} \quad o_t \sim P(o_t|s_t) \qquad (8)$$

---

[1]The full set of SMC samples can instead be used to handle ambiguous sentences with multiple plausible ELoT translations. We leave the study of this to future work.

3.2.1 Modeling Perception and Action

Two crucial aspects of our BToM module are how it models belief updating (Eq. 6) as the result of perception, and how it models goal-directed action given the agent's uncertain beliefs (Eq. 7). To model perception, we represent an agent's belief b_t as a probability-weighted collection $\{(\tilde{s}_i, w_i)\}_{i=1}^{n_s}$ of possible environment states $\tilde{s}_i$ (which are represented in turn as collections of ELoT predicates). Given an observation of the environment s_t (e.g. observing that a box is empty), the agent updates its belief by *filtering* out inconsistent hypotheses $\tilde{s}_i$, setting $w_i = 0$.

As for goal-directed action, our model builds upon methods for epistemic planning (Bolander, 2017) and belief-space planning in POMDPs (Littman et al., 1995). Given a belief b_t, the agent engages in *instrumental planning* to achieve their goal g, which requires achieving instrumental subgoals (e.g. picking up keys), but also gathering goal-relevant information (e.g. finding out if a key is in a certain box). We model this by assuming that the agent acts by approximately minimizing a *cost-to-go estimate* $\hat{Q}_g(b_t, a)$: An estimate of the optimal cost $Q_g^*(b_t, a)$ of reaching g after action a starting from one's (uncertain) belief b_t. Action selection can thus be modeled by a Boltzmann distribution over these $\hat{Q}_g$ estimates:

$$P(a_t|b_t, g) \propto \exp\left(-\beta \hat{Q}_g(b_t, a)\right) \qquad (9)$$

To estimate Q_g^* efficiently, we follow recent advances in inverse planning (Zhi-Xuan et al., 2024b) by computing the Q_{MDP} approximation (Hauskrecht, 2000) of Q_g^*, averaging over the Q-values for each hypothesis $(\tilde{s}_i, w_i)$ in the belief b_t:

$$\hat{Q}_g(b_t, a) = \sum_{(\tilde{s}_i, w_i) \in b_t} w_i \cdot Q_g^*(\tilde{s}_i, a) \qquad (10)$$

The cost-to-go $Q_g^*(s, a)$ from a known state s can itself be efficiently estimated by searching for a shortest path to g from s (Monfort et al., 2015).

3.2.2 Joint Inference of Goals and Beliefs

With this generative model, observers can jointly infer the agent's goal g, belief history $b_{0:T}$, and environment trajectory $s_{0:T}$ given observations of the agent's actions $a_{1:T}$ and partial observations $o_{0:T}$ of the environment:

$$P(g, b_{0:T}, s_{0:T}|a_{1:T}, o_{0:T}) \propto \qquad (11)$$
$$P(g, s_0, b_0) \prod_{t=1}^{T} P(b_t, a_t, s_t, o_t|b_{t-1}, s_{t-1})$$

To ensure tractable posterior inference, we considered only the set of initial states $\mathcal{S}_0$ consistent with the initial observation o_0, and a discrete set $\mathcal{B}_0$ of possible beliefs b_0 sufficient to model comparative likelihood claims (e.g. *"The key is more likely in box 1 than 2."*). Specifically, we consider all beliefs formed by distributing k particles across $n_s := |\mathcal{S}_0|$ states, resulting in $n_b := |\mathcal{B}_0| = \binom{n_s+k-1}{k}$ distributions. We then perform *exact* Bayesian inference over all combinations of goals g, initial beliefs b_0, and states s_0, which we implement as a variant of Sequential Inverse Plan Search (Zhi-Xuan et al., 2020) using the Gen probabilistic programming system (Cusumano-Towner et al., 2019). More algorithmic details are provided in the Appendix.

3.2.3 Evaluating Epistemic Sentences

By inferring the agent's belief history $b_{0:T}$, we can compute the probability $\mathbf{Pr}(A, \phi)$ of a formula ϕ at time t as the expected truth value under b_t:

$$\mathbf{Pr}(A, \phi) = \sum_{(\tilde{s}_i, w_i) \in b_t} w_i \cdot [\![\phi]\!]_s \qquad (12)$$

We can thus evaluate a epistemic formula φ given a belief b_t and environment state s_t by replacing $\mathbf{Pr}$ terms with their values, then determining the truth of the resulting expression in state s. We denote this operation by $[\![\varphi]\!]_{(s_t, b_t)}$.

However, observers do not have access to the true s_t or b_t, only inferences about them. As such, we model human evaluation of a sentence φ as a probabilistic judgment given their inferences:

$$P([\![\varphi]\!]_{(s_t, b_t)} | a_{1:T}, o_{0:T})$$
$$= \mathbb{E}_{s_t, b_t \sim P(s_t, b_t | a_{1:T}, o_{0:T})} \left[[\![\varphi]\!]_{(s_t, b_t)} \right]$$

While these judgements are made after observing actions up to time T, φ may be *retrospectively* evaluated as a description of the agent's beliefs at any $t \in [0, T]$, with $t=0$ and $t=T$ corresponding to initial and current beliefs respectively.

Following Ying et al. (2024b), we also assume that people provide ratings as if they have a uniform prior U_φ about the truth of φ. Under this prior, ratings of φ can be interpreted as a *normalized likelihood* $\bar{L}([\![\varphi]\!]_{(s_t, b_t)} | a_{1:T}, o_{0:T})$, which measures the likelihood of statement φ relative to its negation $\neg\varphi$. In other words, we assume that humans rate an epistemic claim more highly when they have *evidence* for it. With no evidence, $\bar{L}([\![\varphi]\!]_{(s_t, b_t)} | a_{1:T}, o_{0:T}) = 0.5$. Results investigating the importance of this assumption can be found in the Appendix.

4 Experiments

To evaluate our model on a diverse dataset of epistemic language (Table 2), we conducted a two-part human experiment. We first recruited participants to write English sentences describing the current and initial beliefs of a player character as it navigated a gridworld puzzle that required finding keys hidden in boxes. Next, we asked two groups of participants to rate how likely these sentences were to be true given the player's behavior, with one group rating sentences about the player's current beliefs, and the other rating sentences about initial beliefs.

With this data, we evaluated our model by: (i) assessing the **translation accuracy of our ELoT module**, investigating the importance of our ELoT representation (vs. the lowered form) and its impact on grammar-constrained LLM decoding; (ii) testing our **full LaBToM model** in its ability to capture human interpretation of epistemic language, comparing human ratings against LaBToM inferences. Model and experiment code is available at https://osf.io/xq9c4/.

4.1 Scenario Construction

We constructed 20 scenarios in the Doors, Keys, & Gems environment with varied maze designs and item locations (Figure 3). In each scenario, there were 4 goal gems with different shapes (triangle, square, hexagon, circle), some of which were locked behind doors. Scenarios also had 2 to 3 boxes with up to 2 colored keys among them. The player's actions were varied across scenarios to elicit inferences about a diversity of epistemic states, such as ignorance about key locations or false confidence about the location of a key.

4.2 Collecting Epistemic Language

In the first part of the experiment, we recruited 42 US participants via Prolific (mean age: 36.02, SD: 10.1; 16 women, 26 men). Following a tutorial, participants watched 10 scenario animations, with each stopping before the player reached their goal and all relevant keys were revealed. Participants were then asked to write at least two sentences about the player's likely beliefs at the end of the scenario (*current beliefs*), and another two sentences about the player's beliefs at the start of the scenario (*initial beliefs*). To ensure that these sentences focused on beliefs about the environment, we instructed participants to

Factor	Description	Count	
	Examples	*Current*	*Initial*
Possibility	Sentences with modal verbs such as might, could and must.	66 / 241	46 / 228
	The player believes box 1 may contain a blue key or a red key.		
	The player believes if the red key is not in box 2 then it must be in box 3.		
Probability	Sentences with probability expressions such as likely, uncertain, etc.	28 / 241	28 / 228
	The player thought that box 1 was most likely to contain a red key.		
	The player is unsure what color the key in box 2 will be.		
Compositionality	Sentences that embed compound propositions (conjunctions, disjunctions, etc.).	49 / 241	69 / 228
	The player thinks that there's more likely to be a red key in box 1 or 3 than box 2.		
	The player believes that if box 1 does not have a blue key, then box 3 has a blue key.		
Knowledge	Sentences that make knowledge or ignorance claims.	17 / 241	20 / 228
	The player already knows for sure there is no key in box 1 or box 2.		
	The player did not know if box 2 contained a red key.		

Table 2: Overview of our dataset of 464 human-written epistemic sentences, broken down by factors.

describe the player's beliefs about the contents of the boxes. We excluded 8 participants for failing to follow these instructions, and about one third of remaining sentences (see Appendix). This process left us with 241 (228) statements about current (initial) beliefs, which were then annotated with factors by two experimenters. Table 2 illustrates the diversity of language we collected.

4.3 Evaluating Epistemic Language

For the next part of our experiment, we recruited 94 US participants via Prolific to evaluate current belief statements (mean age = 35.7, SD = 12.4, 67 women, 27 men), and another 104 US participants to evaluate initial belief statements (mean age = 35.27, SD = 11.7, 69 women, 33 men, 2 non-binary). Each participant was shown 10 out of 20 scenario animations, and was asked to rate the goals and beliefs of the player at several judgment points during each animation. For goals, participants were shown a checkbox for each gem, and asked to select all gems likely to be the agent's goal. This served as both an attention check and an additional data source for model validation. For beliefs, participants were shown 2 belief statements selected from our dataset of human-written statements, and asked to rate how likely each statement was on a scale from 1 to 7. These ratings were normalized between 0 and 1 for our analysis. We excluded 7 participants from the current belief condition and 5 from the initial belief condition for low outlying scores on the goal inference subtask.

4.3.1 Statement Selection

To ensure that the belief statements evaluated by our participants were (i) diverse and (ii) rated enough times to ensure sufficient statistical

power (88% power at Cohen's $d = 0.8$), we selected 5 statements per scenario (3 plausible, 2 implausible) from our full dataset of statements. The plausible statements were chosen by sampling many sets of 3 statements out of all those written for a scenario, then selecting the set that scored highest on a diversity metric derived from the factors in Table 2 (see Appendix for details). We then manually added 2 more statements that were originally written for other scenarios, and which we evaluated to be implausible descriptions of the target scenario. Participants were shown 2 of these 5 statements at random in each scenario.

4.4 ELoT Parser Evaluation

We evaluated the suitability of ELoT formulas as representations of epistemic sentences by investigating (i) the impact of using ELoT as the translation target (instead of the lowered form, which lacks epistemic operators besides **Pr**) and (ii) the value of the associated ELoT grammar $G_{\mathcal{E}}$ in constraining LLM outputs.

To do so, we created gold translations of the subset of statements selected for human evaluation (125 statements) into ELoT formulas and their lowered forms. We then compared the translations produced by grammar-constrained SMC decoding ($n_\sigma = 10$ samples) of LLaMa 3.1 8B against these gold translations in terms of both strict equality and approximate semantic equivalence (i.e. cases of reasonable alternative translations of the original English statement). As baselines, we used unconstrained self-consistency sampling (Wang et al., 2023) with LLaMa 3.1 8B and the instruction-tuned Gemini Flash 8B, sampling $n_\sigma = 10$ times and taking the majority answer. All methods used the same prompt format (see Appendix), with 36 example translations.

4.5 LaBToM Fitting and Evaluation

We evaluated our LaBToM model on all 20 scenarios, producing normalized likelihood scores for the 5 current and 5 initial belief statements per scenario. We then computed Pearson's r between these scores and average human ratings. We fit our model parameters to maximize r, fitting the belief thresholds $\Theta := (\theta_{\text{believes}}, \theta_{\text{could}}...)$ via coordinate ascent, and the inverse temperature β via grid search (see Appendix for robustness analyses). This produced the fitted values for Θ in Table 1, and $\beta = 2^{3/2}$. For the set of possible initial agent beliefs $\mathcal{B}_0$, we fixed the number of belief particles to $k = 3$ to ensure tractable exact inference.

Alongside this direct comparison with human-provided ratings, we evaluated our model on the full dataset of 469 human-written statements by pairing each statement with either the scenario it was written for (*in-context*) or 1–2 other scenarios with the same map layout but distinct agent trajectories (*out-of-context*). This allowed us to compare each statement's in-context normalized likelihood $\bar{L}$ with its (average) out-of-context likelihood score. Reasonable models of epistemic language interpretation should assign higher likelihood scores to most statements when they are evaluated in-context vs. out-of-context.

4.6 Baselines

To assess the import of a sufficiently rich theory of mind for epistemic language understanding, we evaluated several ablations of our model under simplified assumptions about the agent's beliefs or planning abilities. We also evaluated two state-of-the-art multi-modal LLMs, thereby testing the degree to which grounded evaluation of epistemic language can be achieved with sufficient scale:

True Belief. The True Belief ablation assumes that the observed agent has fully accurate beliefs about the environment (i.e. they already know where all the keys are located), equivalent to the full model from Ying et al. (2024b). The observer starts with a uniform prior over these true beliefs.

Non-Planning. The Non-Planning ablation assumes that the agent is incapable of planning towards instrumental subgoals (such as keys), and instead optimizes the heuristic of moving physically closer to the goal. This is implemented by using the Manhattan distance to the goal as the agent's cost-to-go estimate $\hat{Q}_g$.

Parser	Translation Accuracy (s.e.)			
	ELoT		Lowered	
	Exact	*Equiv.*	*Exact*	*Equiv.*
Grammar-Constrained SMC (Loula et al., 2025)				
LLaMa 3.1 8B	**0.81**	**0.91**	0.34	0.42
	(.03)	(.03)	(.04)	(.04)
Unconstrained Sampling + Majority Vote				
LLaMa 3.1 8B	0.65	0.69	0.32	0.36
	(.04)	(.04)	(.04)	(.04)
Gemini Flash 8B	0.69	0.80	0.60	0.68
	(.04)	(.03)	(.04)	(.04)

Table 3: **Translation accuracy for ELoT vs. lowered formulas as the translation target**, compared across sampling methods and LLMs in terms of exact equality (*Exact*) or semantic equivalence (*Equiv.*). ELoT serves as a better target *and* constraint for translation.

Multi-modal LLMs. We use GPT-4o (text & image input, `gpt-4o-2024-05-13`) and Gemini 1.5 Pro (text & image or video input) as multi-modal LLM baselines, providing them the same instructions as human participants (see Appendix for a unimodal baseline). Each baseline was run 3 times with a temperature of 1.0. In addition to providing an image or video of the scenario showing the actions up to each judgment point, we used the following prompting methods:

Plans: The prompt plainly describes the agent's actions over time (e.g. *the player moves right five times*), and also describes the agent's observations (e.g. *the player opens box 1 and finds a red key*).

Narratives: A rich narrative of agent behavior is included in the prompt, providing key information about the scene (e.g. which gems are locked behind doors, which keys are visible), while also describing the agent's movements in relation to relevant objects (e.g. *the player moves right five times, going past box 1 towards box 2*).

Few-Shot Prompting: After describing the plan or narrative, we provide the LLM with human ratings for the 4 other statements tied that scenario, before querying its rating for the target statement.

5 Results

ELoT is a superior translation target and guide for (grammar-constrained) LLM parsing. We report translation accuracy in Table 3. In confirmation of the idea that our ELoT formalism captures the semantics of belief sentences at the right level of granularity, we find that translating English into ELoT is up to 2.4 times as accurate as translating into the lowered form, even though they have the same expressive capacity. ELoT also increases the benefit of grammar-constrained

	Parser	Example
English		The player believes that box 3 is empty.
ELoT	*Gold*	believes(player, formula(empty(box3)))
	Grammar-Cons.	believes(player, formula(empty(box3)))
	Unconstrained	believes(player, empty(box3)) **(Syntax Error)**
Lowered	*Gold*	>=(prob_of(player, empty(box3)), threshold(believes))
	Grammar-Cons.	and(>=(prob_of(player, empty(box3)), threshold(believes)), empty(box3)) **(Semantic Error)**
	Unconstrained	>=(prob_of(player, empty(box3)), believes) **(Syntax Error)**

Table 4: Example translations from English to ELoT and lowered formulas using LLaMa 3.1 8B.

Model	Human Correlation r (s.e.)		
	All	*Current*	*Initial*
LaBToM			
Full (ours)	**0.76 (0.01)**	**0.78 (0.01)**	**0.72 (0.01)**
Non-Planning	0.40 (0.01)	0.58 (0.01)	0.07 (0.01)
True Belief	0.10 (0.01)	0.09 (0.01)	0.09 (0.02)
GPT-4o (OpenAI et al., 2023)			
I+Na+FS	0.52 (0.01)	0.59 (0.01)	0.41 (0.01)
I+Na	0.48 (0.01)	0.52 (0.01)	0.41 (0.01)
I+Pl	0.28 (0.01)	0.32 (0.01)	0.18 (0.01)
Gemini 1.5 Pro (Gemini Team et al., 2024)			
V+Na+FS	0.23 (0.01)	0.28 (0.01)	0.14 (0.02)
I+Na+FS	0.22 (0.01)	0.29 (0.01)	0.11 (0.02)

I - Image, V - Video, Pl - Plans, Na - Narratives, FS - Few-Shot

Table 5: Model correlations with human ratings of epistemic language. LaBToM correlates strongly with humans, whereas multimodal LLMs struggle to do so.

SMC decoding, leading to an improvement of 0.20 in equivalence accuracy for LLaMa 3.1 8B. In contrast, using the lowered grammar leads to only an 0.06 improvement over unconstrained generation. We show example translations and errors for each representation in Table 4. These findings highlight the importance of "language-of-thought" representations that are aligned with the structure of natural language, in line with work tying language use to the structure and acquisition of compositional concepts (Wong et al., 2022).

Overall, grammar-constrained SMC decoding into ELoT with LLaMa 3.1 8B performs the best, achieving 91% equivalence with our gold translations. Upon inspection (see Appendix), we find that unconstrained sampling often leads to syntax errors not corrected for by majority voting. Gemini Flash 8B also exhibits low sample diversity typical of instruction tuning, such that majority voting gives no improvement. In comparison, grammar-constrained SMC produces not just accurate translations, but a *distribution* over reasonable translations, since SMC performs posterior sampling rather than optimization (Loula et al., 2025; Lew et al., 2023b). In future work, this could be used to more precisely model the interpretation of ambiguous epistemic language.
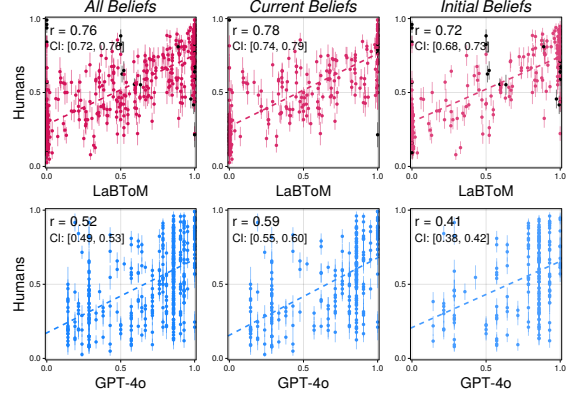

Figure 2: Human correlation plots for LaBToM and GPT-4o. Black dots are from statements not accurately translated to ELoT. LaBToM provides a much stronger qualitative fit with human ratings compared to the best GPT-4o baseline, which fails to use the full scale.

LaBToM correlates highly with human ratings of epistemic statements. As Table 5 shows, our full model produces statement scores that correlate highly with human ratings ($r = 0.76$) across both current and initial belief conditions. In the Appendix, we show results using the gold ELoT translations ($r = 0.81$), and a per-factor breakdown. Plotting average human judgments against LaBToM inferences (Figure 2, Row 1), we also find a strong qualitative fit: Factoring out incorrect ELoT translations (black dots), LaBToM generally assigns high or low ratings to sentences when humans do, using the ends of the scale as appropriate. In contrast, our ablated models do poorly, either because they fail to track how the player updates their beliefs (True Belief, Current $r = 0.09$), or fail to infer the player's initial beliefs (Non-Planning, Initial $r = 0.07$). These findings hold even when the parameters of our full model are adversarially optimized (see Appendix).

SotA multimodal LLMs struggle at grounded evaluation of epistemic language. The multimodal LLM baselines also perform less well than LaBToM despite extensive prompting, with the
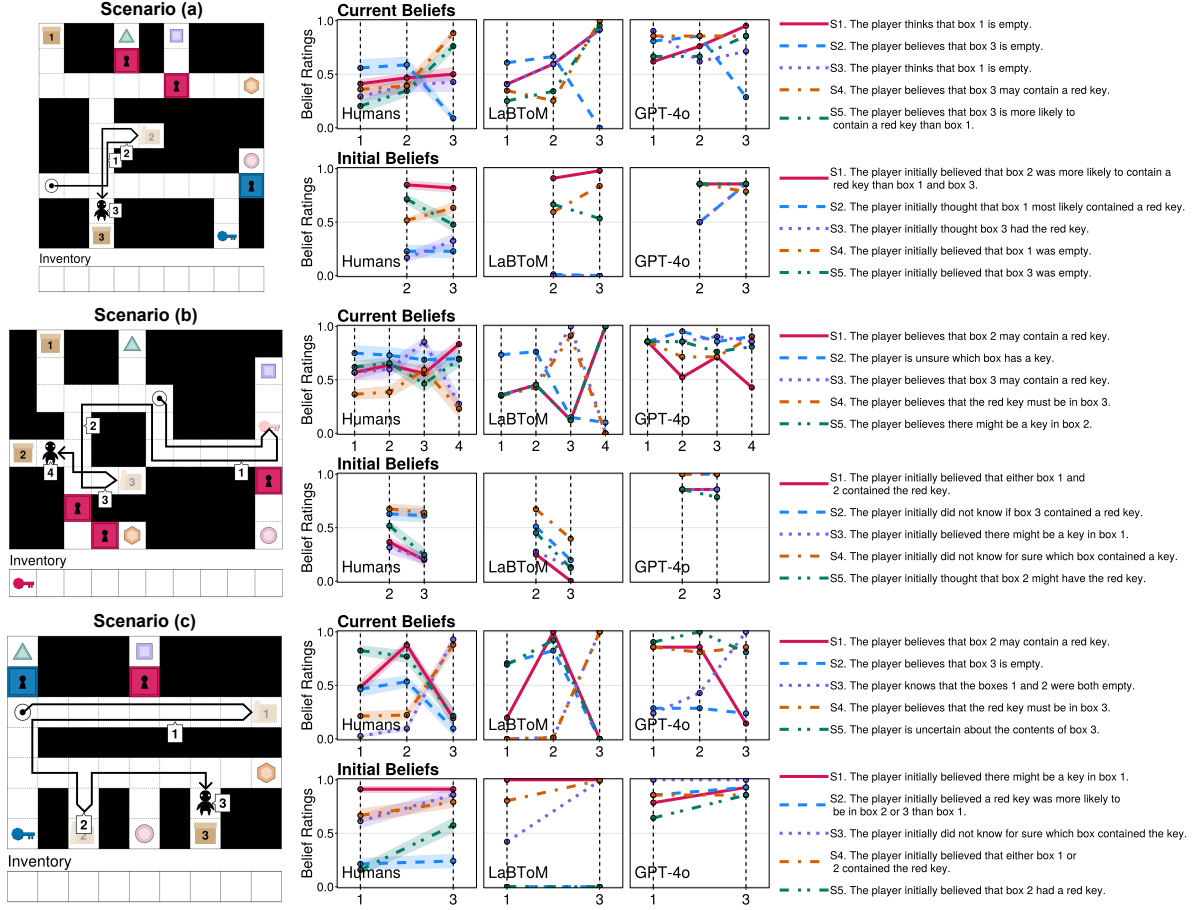
Figure 3: Step-by-step ratings by humans and models across three scenarios. Judgement points are annotated on each map, and show the player's location *before* opening the nearest box. Keys picked up along the way are shown in light colors. Our model largely matches human responses qualitatively and quantitatively, unlike GPT-4o.

strongest LLM baseline (GPT-4o with images, narratives, and few-shot prompting) achieving human correlations of only 0.59 and 0.41 respectively. Increasing information in the prompt improves performance, although the Gemini models do poorly regardless, even with full scenario videos. We find that LLM performance is lower on initial belief sentences, suggesting that they are better at tracking how agents' current beliefs change, but worse at inferring past beliefs.

Figure 2 illustrates the difference between our model and the best performing GPT-4o baseline (I+Na+FS) in greater detail. Despite few-shot prompting, GPT-4o tends to assign ratings of 0.8 or more to many statements that humans find quite unlikely, suggesting that LLMs struggle to account for evidence *against* a belief claim.

LaBToM captures how human evaluations of epistemic language change with agent behavior. Our model predicts that human evaluations of epistemic sentences should change systematically

as they gain more information about an agent's percepts and actions. As Figure 3 illustrates, this is what we find. When the player sees that a box is empty, both humans and our model sharply decrease their ratings for statements claiming that the player believes a key is in that box (e.g. Fig. 3b, S4 Current). When the player approaches one box instead of another, both humans and LaBToM gain confidence in statements about the *relative likelihood* of key locations (Fig. 3a, S5 Current).

We analyze just one scenario (Figure 3a) in detail. Here, the player first walks upwards away from box 3 (Judgment Point 1), leading both humans and LaBToM to assign scores of greater than 0.5 to *"The player believes that box 3 is empty"* and *"The player initially believed that box 3 was empty"*. In contrast, the modal sentence *"The player believes that box 3 may contain a red key"* is rated lower. This is because if the player did believe that box 3 might have the (red) key, then it is more likely that they would have looked in box 3. However, this does not occur.

After the player opens box 2 and finds it empty, then walks back down towards box 3 (Judgment Point 3), both humans and our model decrease their confidence in the statement *"The player believes that box 3 is empty"*. They also decrease confidence in *"The player initially believed that box 3 was empty"*, but *less sharply*. This is because there are at least two possibilities consistent with the player's actions: They could have firmly believed that box 3 was empty, or they could have just believed that box 3 was *less likely* to contain the relevant key than box 2, without all-out believing that it was empty. This ability to understand beliefs about relative likelihood is made explicit by how ratings change over time for *"The player believes that box 3 is more likely to contain a red key than box 1."* Consistent with the principle of rational action, both humans and LaBToM assign a high score to this sentence once the player goes back to box 3, forgoing box 1.

Across a range of other settings (Figure 3b–c), our model largely captures fine-grained changes in how people evaluate a variety of epistemic expressions, including modal sentences, ignorance claims, and expressions of uncertainty. Unlike prior BToM models that lack language-like belief representations (Baker et al., 2017), LaBToM also distinguishes observer and agent uncertainty, assigning high ratings to claims that the agent is uncertain (Figure 3c, S5 Cur.), and vice versa. In contrast, the best LLM baseline (GPT-4o, I+Na+FS) often fails to adjust its ratings in the same direction as humans, while rating implausible statements highly. We discuss these results more in the Appendix, alongside cases where our model comes apart from humans.

LaBToM distinguishes in-context and out-of-context epistemic language. To investigate how our model generalizes to a larger set of epistemic expressions, we performed the in-context vs. out-of-context likelihood comparison described in Section 4.5 for our full dataset of 469 sentences. We tested the full LaBToM model, ablations and the best applicable LLM baseline from Table 5 (GPT-4o, I+Na). Results are shown in Table 6. For both current and initial belief statements, we find that LaBToM assigns significantly higher scores when a statement is evaluated in-context vs. out-of-context, correctly classifying the context about 70% of the time by assigning a strictly higher in-context score. On closer inspection,

Model	Statement Likelihoods In-Ctx / Out-of-Ctx / Diff.	Accuracy In vs. Out
Current Beliefs (241 statements)		
LaBToM (ours)	**0.78 / 0.48 / +0.30 (0.03)**	**0.71 (0.03)**
Non-Planning	0.70 / 0.54 / +0.16 (0.02)	0.65 (0.03)
True Belief	0.35 / 0.35 / +0.01 (0.01)	0.11 (0.02)
GPT-4o (I+Na)	0.80 / 0.62 / +0.18 (0.03)	0.59 (0.03)
Initial Beliefs (228 statements)		
LaBToM (ours)	**0.73 / 0.51 / +0.22 (0.03)**	**0.70 (0.03)**
Non-Planning	0.70 / 0.70 / -0.01 (0.01)	0.39 (0.03)
True Belief	0.51 / 0.50 / +0.00 (0.01)	0.16 (0.02)
GPT-4o (I+Na)	0.76 / 0.68 / +0.07 (0.02)	0.38 (0.03)

Table 6: **In vs. out-of-context statement evaluation.** LaBToM most accurately distinguishes when epistemic language is evaluated in vs. out-of-context, assigning significantly higher scores in-context (s.e. in brackets).

many statements that are incorrectly classified turn out to be plausible in either context, resulting in equal or close-to-equal scores. The ablated models and GPT-4o perform significantly worse than LaBToM, especially for initial beliefs.

6 Discussion

Our experiments show that, similar to humans, our LaBToM model is able to coherently interpret and adjust its evaluations of natural language statements about agents' beliefs, whereas state-of-the-art multimodal LLMs struggle with this task. This ability is mediated by our ELoT representation and semantic parser: Without a compositional representation of epistemic concepts that is aligned with the structure of natural epistemic language, we find that translation accuracy drops significantly, even when using a grammar-constrained LLM parser. We also find that LaBToM largely distinguishes in vs. out-of-context sentence usage on a large and diverse set of crowd-sourced epistemic language, indicating the generalizability of our approach.

That said, our model is not without limitations. As we discuss at greater length in the Appendix, LaBToM's outputs depart from human judgments in several interesting ways, suggesting the need to account for (i) contextual adaptation of probability thresholds via pragmatic reasoning (Rudin, 2016; Schuster and Degen, 2020), (ii) the role of justification in human's intuitive evaluation of knowledge claims (Alston, 1989), and (iii) bounded human reasoning about logical implications (Smets and Solaki, 2018).

LaBToM is also an ideal observer model that does not scale readily to large belief spaces, leaving open how humans *tractably* infer and evaluate claims about others' beliefs (Van Rooij, 2008), perhaps by focusing on *occurent* beliefs (Bartlett, 2018) that are relevant to others' goals. Finally, LaBToM is a model of how people *interpret* epistemic language, but full understanding also includes the ability to *produce* such language. This could potentially be achieved by inverting the ELoT module of our model, using it to translate salient or conversationally-relevant inferences about an agent's beliefs into natural language. By extending our model in this way, we stand to gain an even richer account of what it means to understand epistemic language.

Acknowledgments

References

Jerry A. Fodor. 1981. Propositional attitudes. In *The Language and Thought Series*, pages 45–63. Harvard University Press.

William P Alston. 1989. *Epistemic justification: Essays in the theory of knowledge*. Cornell University Press.

Yoav Artzi and Luke Zettlemoyer. 2013. Weakly supervised learning of semantic parsers for mapping instructions to actions. *Transactions of the association for computational linguistics*, 1:49–62.

Chris L Baker, Julian Jara-Ettinger, Rebecca Saxe, and Joshua B Tenenbaum. 2017. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1(4):1–10.

Gary Bartlett. 2018. Occurrent states. *Canadian Journal of Philosophy*, 48(1):1–17.

Marcel Binz and Eric Schulz. 2023. Turning large language models into cognitive models. *arXiv preprint arXiv:2306.03917*.

Ned Block. 1987. Advertisement for a semantics for psychology. *Midwest studies in philosophy*, 10(1):615–678.

Thomas Bolander. 2017. A gentle introduction to epistemic planning: The del approach. *arXiv preprint arXiv:1703.02192*.

Jennifer Coates. 1987. Epistemic modality and spoken discourse. *Transactions of the Philological society*, 85(1):110–131.

Sholei Croom, Hanbei Zhou, and Chaz Firestone. 2023. Seeing and understanding epistemic actions. *Proceedings of the National Academy of Sciences*, 120(49):e2303162120.

Marco F Cusumano-Towner, Feras A Saad, Alexander K Lew, and Vikash K Mansinghka. 2019. Gen: a general-purpose probabilistic programming system with programmable inference. In *Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation*, pages 221–236. ACM.

Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. 2006. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(3):411–436.

Andy Egan and Brian Weatherson. 2011. *Epistemic modality*. Oxford University Press.

J Fodor. 1987. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. MIT Press.

Jerry A Fodor et al. 1975. *The language of thought*, volume 5. Harvard university press Cambridge, MA.

Gottlob Frege. 1948. Sense and reference. *The Philosophical Review*, 57(3):209–230.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut,

Johan Schalkwyk, Andrew M. Dai, Anja Hauth, et al. 2024. Gemini: A family of highly capable multimodal models.

Noah D Goodman and Daniel Lassiter. 2015. Probabilistic semantics and pragmatics uncertainty in language and thought. *The handbook of contemporary semantic theory*, pages 655–686.

Udo Hahn and Christine Engelmann. 2014. Grounding epistemic modality in speakers' judgments. In *Pacific Rim International Conference on Artificial Intelligence*, pages 654–667. Springer.

Gilbert Harman. 1982. Conceptual role semantics. *Notre Dame Journal of Formal Logic*, 23(2):242–256.

Milos Hauskrecht. 2000. Value-function approximations for partially observable markov decision processes. *Journal of Artificial Intelligence Research*, 13:33–94.

Kaarlo Jaakko Juhani Hintikka. 1962. *Knowledge and Belief: An Introduction to the Logic of the Two Notions*. Cornell University Press, Ithaca, NY, USA.

Ray Jackendoff. 2003. Précis of foundations of language: Brain, meaning, grammar, evolution. *Behavioral and brain sciences*, 26(6):651–665.

Julian Jara-Ettinger, Laura E Schulz, and Joshua B Tenenbaum. 2020. The naive utility calculus as a unified, quantitative framework for action understanding. *Cognitive Psychology*, 123:101334.

Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer Ullman, Antonio Torralba, Joshua Tenenbaum, and Tianmin Shu. 2023. MMToM-QA: Multimodal theory of mind question answering. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*.

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

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023. Fantom: A benchmark for stress-testing machine theory of mind in interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14397–14413.

Brenden M Lake and Gregory L Murphy. 2023. Word meaning in minds and machines. *Psychological review*, 130(2):401.

George Lakoff. 1988. Cognitive semantics. *Meaning and Mental Representation*.

Daniel Lassiter. 2010. Gradable epistemic modals, probability, and scale structure. In *Semantics and Linguistic Theory*, pages 197–215.

Daniel Lassiter. 2017. *Graded modality: Qualitative and quantitative perspectives*. Oxford University Press.

Alexander K Lew, George Matheos, Tan Zhi-Xuan, Matin Ghavamizadeh, Nishad Gothoskar, Stuart Russell, and Vikash K Mansinghka. 2023a. SMCP3: Sequential Monte Carlo with probabilistic program proposals. In *International conference on artificial intelligence and statistics*, pages 7061–7088. PMLR.

Alexander K Lew, Tan Zhi-Xuan, Gabriel Grand, and Vikash Mansinghka. 2023b. Sequential monte carlo steering of large language models using probabilistic programs. In *ICML 2023 Workshop: Sampling and Optimization in Discrete Space*.

Michael L Littman, Anthony R Cassandra, and Leslie Pack Kaelbling. 1995. Learning policies for partially observable environments: Scaling up. In *Machine Learning Proceedings 1995*, pages 362–370. Elsevier.

Brian Loar. 1981. *Mind and Meaning*. CUP Archive.

João Loula, Benjamin LeBrun, Li Du, Ben Lipkin, Clemente Pasti, Gabriel Grand, Tianyu Liu, Yahya Emara, Marjorie Freedman, Jason Eisner, Ryan Cotterell, Vikash Mansinghka, Alexander K. Lew, Tim Vieira, and Timothy J. O'Donnell. 2025. Syntactic and semantic

control of large language models via sequential monte carlo. In *The Thirteenth International Conference on Learning Representations*.

John MacFarlane. 2011. Are assessment-sensitive. *Epistemic modality*, pages 144–179.

Drew McDermott, Malik Ghallab, Adele Howe, Craig Knoblock, Ashwin Ram, Manuela Veloso, Daniel Weld, and David Wilkins. 1998. PDDL - the Planning Domain Definition Language.

Björn Meder, Ralf Mayrhofer, and Azzurra Ruggeri. 2022. Developmental trajectories in the understanding of everyday uncertainty terms. *Topics in Cognitive Science*, 14(2):258–281.

Hugo Mercier and Dan Sperber. 2017. *The Enigma of Reason*. Harvard University Press.

Mathew Monfort, Brenden M Lake, Brian Ziebart, Patrick Lucey, and Josh Tenenbaum. 2015. Softstar: Heuristic-guided probabilistic inference. *Advances in Neural Information Processing Systems*, 28.

Sarah Moss. 2015. On the semantics and pragmatics of epistemic vocabulary. *Semantics and Pragmatics*, 8:5–1.

Laura O'Mahony, Leo Grinsztajn, Hailey Schoelkopf, and Stella Biderman. 2024. Attributing mode collapse in the fine-tuning of large language models. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report.

Barbara Hall Partee. 1973. The semantics of belief-sentences. In *Approaches to natural language: Proceedings of the 1970 Stanford workshop on grammar and semantics*, pages 309–336. Springer.

Steven Piantadosi and Felix Hill. 2022. Meaning without reference in large language models. In *NeurIPS 2022 Workshop on Neuro Causal and Symbolic AI (nCSI)*.

Steven Thomas Piantadosi. 2011. *Learning and the Language of Thought*. Ph.D. thesis, Massachusetts Institute of Technology.

Deniz Rudin. 2016. Deriving a variable-strength might. In *Proceedings of Sinn und Bedeutung*, volume 20, pages 587–603.

Bertrand Russell. 1903. *Principles of mathematics*. Routledge.

Stephen Schiffer. 2003. *The Things We Mean*. Clarendon Press.

Torsten Scholak, Nathan Schucher, and Dzmitry Bahdanau. 2021. PICARD: Parsing incrementally for constrained auto-regressive decoding from language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9895–9901, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sebastian Schuster and Judith Degen. 2020. I know what you're probably going to say: Listener adaptation to variable use of uncertainty expressions. *Cognition*, 203:104285.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2024. 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)*, pages 2257–2273.

Richard Shin, Christopher Lin, Sam Thomson, Charles Chen Jr, Subhro Roy, Emmanouil Antonios Platanios, Adam Pauls, Dan Klein, Jason Eisner, and Benjamin Van Durme. 2021. Constrained language models yield few-shot semantic parsers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7699–7715.

Maayan Shvo, Toryn Q Klassen, Shirin Sohrabi, and Sheila A McIlraith. 2020. Epistemic plan recognition. In *Proceedings of the 19th International Conference on Autonomous*

Agents and MultiAgent Systems, pages 1251–1259.

Sonja Smets and Anthia Solaki. 2018. The effort of reasoning: Modelling the inference steps of boundedly rational agents. In *Logic, Language, Information, and Computation: 25th International Workshop, WoLLIC 2018, Bogota, Colombia, July 24-27, 2018, Proceedings 25*, pages 307–324. Springer.

James WA Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, et al. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, pages 1–11.

Iris Van Rooij. 2008. The tractable cognition thesis. *Cognitive Science*, 32(6):939–984.

Kai Von Fintel and Irene Heim. 2011. Intensional semantics. *Unpublished lecture notes*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Caroline J Wesson and Briony D Pulford. 2009. Verbal expressions of confidence and doubt. *Psychological Reports*, 105(1):151–160.

Brandon T Willard and Rémi Louf. 2023. Efficient guided generation for large language models. *arXiv preprint arXiv:2307.09702*.

Lionel Wong, Gabriel Grand, Alexander K Lew, Noah D Goodman, Vikash K Mansinghka, Jacob Andreas, and Joshua B Tenenbaum. 2023. From word models to world models: Translating from natural language to the probabilistic language of thought. *arXiv preprint arXiv:2306.12672*.

Lionel Wong, William P McCarthy, Gabriel Grand, Yoni Friedman, Josh Tenenbaum, Jacob Andreas, Robert Hawkins, and Judith E Fan. 2022. Identifying concept libraries from language about object structure. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

Seth Yalcin. 2007. Epistemic modals. *Mind*, 116(464):983–1026.

Lance Ying, Katherine M Collins, Megan Wei, Cedegao E Zhang, Tan Zhi-Xuan, Adrian Weller, Joshua B Tenenbaum, and Lionel Wong. 2023a. The Neuro-Symbolic Inverse Planning Engine (NIPE): Modeling probabilistic social inferences from linguistic inputs. *arXiv preprint arXiv:2306.14325*.

Lance Ying, Kunal Jha, Shivam Aarya, Joshua B Tenenbaum, Antonio Torralba, and Tianmin Shu. 2024a. GOMA: Proactive embodied cooperative communication via goal-oriented mental alignment. *arXiv preprint arXiv:2403.11075*.

Lance Ying, Tan Zhi-Xuan, Vikash Mansinghka, and Joshua B Tenenbaum. 2023b. Inferring the goals of communicating agents from actions and instructions. In *Proceedings of the AAAI Symposium Series*, volume 2, pages 26–33.

Lance Ying, Tan Zhi-Xuan, Lionel Wong, Vikash Mansinghka, and Josh Tenenbaum. 2024b. Grounding language about belief in a Bayesian theory-of-mind. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

Cedegao E Zhang, Lionel Wong, Gabriel Grand, and Joshua B Tenenbaum. 2023. Grounded physical language understanding with probabilistic programs and simulated worlds. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 45.

Tan Zhi-Xuan, Nishad Gothoskar, Falk Pollok, Dan Gutfreund, Joshua B Tenenbaum, and Vikash K Mansinghka. 2022. Solving the Baby Intuitions Benchmark with a Hierarchically Bayesian Theory of Mind. In *RSS 2022 Workshop on Social Intelligence in Humans and Robots*.

Tan Zhi-Xuan, Gloria Kang, Vikash Mansinghka, and Joshua B Tenenbaum. 2024a. Infinite ends from finite samples: Open-ended goal inference as top-down Bayesian filtering of bottom-up proposals. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 46(46).

Tan Zhi-Xuan, Jordyn Mann, Tom Silver, Josh Tenenbaum, and Vikash Mansinghka. 2020.

Online Bayesian goal inference for boundedly rational planning agents. *Advances in Neural Information Processing Systems*, 33.

Tan Zhi-Xuan, Lance Ying, Vikash Mansinghka, and Joshua B Tenenbaum. 2024b. Pragmatic instruction following and goal assistance via cooperative language guided inverse plan search. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*.

A Dataset Collection

A.1 Experimental Procedure

The interface used for collecting the statements and evaluating statements are shown in Figure A1. Participants first completed a tutorial that explained the task and experimental interface, then answered 5 comprehension questions before proceeding to the main experiment. In the main experiment, they were shown 10 out of the 20 stimuli in a randomized order.

To incentivize accurate but calibrated responses, participants were rewarded for accurately guessing the true goal. Specifically, they earned $1/N$ bonus points if they selected N goals out of which one was the true goal, but 0 points if none of their selected goals was the true goal. Participants were paid US\$1 for every 40 bonus points they earned, on top of a base pay of US\$15/hr.

Figure A1: Interfaces used for collecting (top) and evaluating (bottom) epistemic language.

A.2 Statement Post-Processing and Annotation

Once statements were collected, two experimenters independently annotated whether each statement was valid for inclusion. We excluded invalid sentences based on two criteria: (i) whether the statement had the right tense (present for current beliefs, past for initial beliefs) and (ii) whether the statement referred to beliefs about the boxes. We also corrected minor grammatical errors and normalized statements to the form *"The player + [believes/knows/thinks/expects/is sure/is uncertain, etc.]..."* for current beliefs and *"The player initially + [believed/knew/thought/expected/was sure/uncertain, etc.]..."* for initial beliefs.

After filtering and normalization, two experimenters independently annotated the statements based on four factors: possibility, probability, compositionality, and knowledge. These factors are not mutually exclusive, so a statement could be annotated with any combination of factors. The codes for possibility were [*may, might, can, could, should, must, none*] and the codes for probability were [*certain, uncertain, likely, unlikely, none*]. The codes for compositionality and knowledge were binary [0, 1]. The annotators agreed on 95% of the codes and discussed to resolve their differences.

A.3 Selecting Diverse Statements for Human Evaluation

After annotation, we selected a set of 3 plausible and 2 implausible statements per scenario for evaluation by human raters. Implausible statements were manually selected from other scenarios to be implausible in their scenario of evaluation. To select the 3 plausible statements, we sampled 100 subsets of 3 statements out of all statements written for that scenario, then computed a diversity score for each set S:

$$\text{Score}(S) = \frac{1}{4}|S_{\text{possibility}}| + \frac{1}{4}|S_{\text{probability}}| + \frac{1}{4}|S_{\text{compositionality}}| + \frac{1}{4}|S_{\text{knowledge}}| \tag{13}$$

where $|S_{\text{possibility}}|$ indicates the number of unique possibility codes in set S and vice versa. We then chose the set with the highest diversity score among the 100 sampled sets.

B Model Configuration

B.1 Belief-Space Sequential Inverse Plan Search (BSIPS)

Our BToM inference algorithm is a belief-space policy variant of Sequential Inverse Plan Search (SIPS, Zhi-Xuan et al. (2020)), which uses policies to evaluate action likelihoods as in recent extensions of SIPS (Zhi-Xuan et al., 2024b,a). Perhaps surprisingly, Belief-Space SIPS (BSIPS) is able to *exactly* compute the posterior over beliefs, goals, and states (Equation 11) without any Monte Carlo approximation: For the scenarios we considered, there were between 120 and 5940 possible combinations of goals, initial states, and beliefs. Enumerative inference over these hypotheses could run as fast as 0.1s/action (120 hypotheses), going up to 20s/action (5490 hypotheses). Experiments were conducted with a i7-1370P 1.90 GHz CPU and 64 GB RAM. Code can be found at `https://osf.io/xq9c4/`.

Algorithm 1 provides the pseudocode for BSIPS. At each step t, we simulate how the environment changes, and how the agent updates their beliefs based on what they see (L6-7). Next, we efficiently compute belief-space Q-values by leveraging the Q_{MDP} approximation described in Section 3.2.1. This involves averaging over Q-values for each state hypothesis $\tilde{s}$ in the agent's belief b_t^i, which can be done cheaply by memoizing and reusing shortest-path computations across all belief hypotheses (L9-13). The Q-values allow us to compute the likelihood of the observed action a_t, allowing us to reweight each hypothesis by how well it explains the observations (L14-16)

For epistemic language evaluation at this scale, the technical challenges are mostly representational, not algorithmic. However, scaling to larger spaces of goals (Zhi-Xuan et al., 2024a) and (belief) states will require additional layers of Monte Carlo approximation. Our implementation in Gen (Cusumano-Towner et al., 2019) can naturally be extended to these cases, e.g. by leveraging Sequential Monte Carlo for approximate inference of initial states (Del Moral et al., 2006; Lew et al., 2023a).

Algorithm 1 Belief-Space Sequential Inverse Plan Search (BSIPS)

1: **procedure** BSIPS($\mathcal{G}, \mathcal{S}_0, \mathcal{B}_0, a_{1:T}, o_{0:T}$)
2:      $\mathcal{H} \leftarrow \mathcal{G} \times \mathcal{S}_0 \times \mathcal{B}_0$ ▷ Enumerate all hypotheses (goal, belief & state combinations).
3:      $\mathcal{W} \leftarrow \{w^i := P(o_0|s_0^i)\}_{i=1}^{|\mathcal{H}|}$ ▷ Initialize (unnormalized) weights for all hypotheses.
4:      **for** $t \in [1, T]$ **do**
5:          **for** $h^i := (g^i, s_{0:t-1}^i, b_{0:t-1}^i) \in \mathcal{H}$ **do**
6:              $s_t^i \leftarrow$ STATE-TRANSITION(s_{t-1}^i, a_{t-1}) ▷ Simulate next environment state.
7:              $b_t^i \leftarrow$ BELIEF-UPDATE$(b_{t-1}^i, s_t^i, a_{t-1})$ ▷ Simulate agent's belief update.
8:              $h^i \leftarrow (g^i, s_{0:t}^i, b_{0:t}^i)$
9:              $Q_{\text{Bel}}(g^i, b_t^i, \tilde{a}) \leftarrow 0$ **for** $\tilde{a} \in$ VALID-ACTIONS(b_t^i) ▷ Initialize belief-space Q-values.
10:              **for** $(\tilde{s}, \tilde{w}) \in b_t^i$ **and** $\tilde{a} \in$ VALID-ACTIONS$(\tilde{s})$ **do**   ▷ Iterate over environment states in agent's belief.
11:                  $Q^*(g^i, \tilde{s}, \tilde{a}) \leftarrow$ MEMOIZED(PATH-COST$(\tilde{s}, \tilde{a}, g^i)$)   ▷ Compute shortest path cost to goal (memoized).
12:                  $Q_{\text{Bel}}(g^i, b_t^i, \tilde{a}) \leftarrow Q^*(g^i, b_t^i, \tilde{a}) + \tilde{w} \cdot Q(g^i, \tilde{s}, \tilde{a})$ ▷ Update belief-space Q-values.
13:              **end for**
14:              $P(a_t|b_t^i, g^i) \leftarrow \exp(-\beta Q_{\text{Bel}}(g^i, b_t^i, a_t))/\sum_a \exp(-\beta Q_{\text{Bel}}(g^i, b_t^i, a))$ ▷ Compute likelihood of action a_t.
15:              $w^i \leftarrow w^i \cdot P(a_t|b_t^i, g^i)$ ▷ Update weight with action likelihood.
16:              $w^i \leftarrow w^i \cdot P(o_t|s_t^i)$ ▷ Update weight with observation likelihood.
17:          **end for**
18:      **end for**
19:      **return** $(\mathcal{H}, \mathcal{W})$ ▷ Return all hypotheses and their (unnormalized) weights.
20: **end procedure**

B.2 Parameter Fitting and Robustness Analyses

We performed a grid search over parameters for the LaBToM model. The range of inverse temperatures β for the Boltzmann policy went from 0.5 to 4 in multiplicative increments of $\sqrt{2}$. This produced human correlations between $r = 0.75$ (at $\beta = 0.5$) and $r = 0.81$ (at $\beta = 4$) for current beliefs, and between $r = 0.64$ (at $\beta = 0.5$) and $r = 0.80$ (at $\beta = 2^{3/2}$), with $\beta = 2^{3/2}$ producing the best fit overall. Across all values of β, we found that the full model outperformed the ablated baselines.

We also fitted the threshold parameters Θ used in our ELoT representation. We performed grid-based coordinate ascent with a step size of 0.05, starting from values derived from the literature mapping modal words to probabilities (Wesson and Pulford, 2009; Hahn and Engelmann, 2014; Meder et al., 2022), and limiting the search to a range of 0.2 above and below this starting point. Our initial and final threshold parameters are shown in Table B1. To evaluate threshold sensitivity, we also ran the same procedure to *minimize* correlation with humans. This produced a value of $r = 0.71$, which was still much higher than the next best model (GPT-4o I+Na+FS, $r = 0.52$). Optimizing these thresholds for the True Belief and Non-Planning ablations led to maximal correlations of $r = 0.12$ and $r = 0.47$ respectively.

| | | | | | Thresholds θ | | | | | |
	believes	certain	uncertain	likely	unlikely	could	might	may	should	must
Initial	0.75	0.95	0.50	0.60	0.40	0.20	0.20	0.30	0.80	0.95
Fitted	0.75	0.95	0.70	0.70	0.40	0.20	0.20	0.30	0.80	0.95

Table B1: ELoT probability thresholds before and after fitting.

B.3 ELoT Translation Prompt

For both grammar-constrained SMC and unconstrained sampling , we prompted the LLM with 36 examples to translate natural language statements into ELoT formulas. ELoT formulas were represented in a Prolog-like syntax analogous to the mathematical syntax we show in Table 1. The prompt used is shown below. We only show 10 out of 36 examples due to space constraints, and provide the full set in our code release. For translation of initial belief statements, we use a separate prompt to first translate the sentence from past tense to present tense, then translate the present tense sentence to ELoT.

```
Please translate the statement below into logical form. Here are some examples of statements
  and their translations:

Input: The player knows that box 2 and box 3 are empty.
Output: knows_that(player, formula(and(empty(box2), empty(box3))))
Input: The player knows the color of the keys in all of the boxes.
Output: forall(box(B), knows_about(player, color(C), exists(and(key(K), inside(K, B)), iscolor
  (K, C))))
Input: The player doesn't know that there is a blue key in box 2.
Output: not_knows_that(player, formula(exists(and(key(K), iscolor(K, red)), inside(K, box2))))
Input: The player is sure of the color of the key in box 4.
Output: certain_about(player, color(C), exists(and(key(K), inside(K, box4)), iscolor(K, C)))
Input: The player is uncertain about what's in box 2.
Output: uncertain_about(player, color(C), exists(and(key(K), inside(K, box2)), iscolor(K, C)))
Input: The player believes that there is a key in box 4.
Output: believes(player, formula(exists(key(K), inside(K, box4))))
Input: The player thinks that there is a red key in either box 1 or box 3.
Output: believes(player, formula(exists(and(key(K), iscolor(K, red)), or(inside(K, box1),
  inside(K, box3)))))
Input: The player thinks there might be a key in box 1 or box 2.
Output: believes(player, might(exists(key(K), or(inside(K, box1), inside(K, box2)))))
Input: The player thinks there is likely a key in box 2.
Output: believes(player, likely(exists(key(K), inside(K, box2))))
```

B.4 LLM Baseline Prompts

Below we show the prompts we used for our multimodal LLM baselines (GPT-4o and Gemini 1.5 Pro). Associated images and videos can be found in our code and dataset release.

```
[IMAGE OR VIDEO]

You're watching someone play the treasure game shown above.

The player controls a character, and their goal is collect one of the four gems (triangle,
  square, hexagon, or circle).

The rules of the game are as follows:
  - The player can move on the white squares.
  - The player has a full view of the map at all time.
  - The player's goal is to collect exactly one target gem.
  - Keys unlock doors of the same color (e.g. red keys unlock red doors).
  - Each key can only be used once. Keys disappear after use.
  - Each box may be empty or contain exactly one key.
  - The player may or may not know what's in each box.
  - Neither you nor the player can see what's hidden in each box. But both of you can see all
    other objects in the scene.
  - There are at most two keys hidden among the boxes.
  - The player knows that the puzzle is solvable, which means there are just enough keys to
    reach any of the target gems.
  - The keys and doors are labeled. The labels are shown on the top right corner of each cell.
  - Your task is to figure out what the player's goal is, and also what the player initially
    believed about the contents of the boxes.

Now you observe the following:

[INSERT PLAN OR NARRATIVE]

Given this information, which gem(s) are most likely to be the human agent's goal? And how
  would you rate the following statement about the player's current belief from 1 (definitely
  false) to 7 (definitely true)? Rate 4 if you think there is an equal chance of the statement
   being true and false.

Please rate the following statement:

[INSERT STATEMENT]

Please respond in the following JSON format, indicating all gems that you think are likely to
  be the human's goal, and your rating as a number from 1 to 7.

{
goal: [gems...],
rating: x ,
}

The gems should be any of [triangle, square, hexagon, circle] and you can indicate all the
  likely goal gems in your response. The rating should be an integer from 1 to 7. Please
  provide an explanation to your response.
```

An example of plan-based prompting:

```
The player moves right three times, then down twice, and finally left.
```

An example of narrative-based prompting, which provides more contextual information:

```
The square gem is locked behind a red door. The triangle, circle and hexagon gems are not
  locked behind any doors. There are three boxes. No keys are visible in the scene.
The player moves right three times, then down twice, and finally left towards box 2 and away
  from box 3.
```

For few shot prompting, the following text is added before showing the statement to be evaluated. We
provide 4 examples of human ratings:

```
Here is how other people have rated other statements for this stimulus:

{% for statement in statements %}
statement {{ loop.index }}: {{ statement.text }}
rating: {{ statement.rating }}

{% end for %}
```

C Additional Results and Experiment Details

C.1 ELoT Parser Evaluation

When evaluating translations, we performed both an exact equality check against our gold translations, and also a manual check for approximate semantic equivalence (i.e. cases where the parser gave a reasonable alternative translation of the original English statement). For replicability, we include these manual equivalence annotations in our code release.

Table C2 shows characteristic ELoT translation errors produced by each parsing method. Many of the the strict errors from grammar-

LLaMa 3.1 8B + Grammar-Constrained SMC	
Statement	The player initially expected to find a key in box 3.
Gold Trans.	believes(player, formula(exists(key(K), inside(K, box3))))
Translation	(67%) believes(player, likely(exists(key(K), inside(K, box3))))
	(32%) believes(player, formula(exists(key(K), inside(K, box3))))
Error	"Expected" is interpreted as "likely" instead of straightforward belief.
LLaMa 3.1 8B Unconstrained + Majority Vote	
Statement	The player initially thought that box 2 contained a red key.
Gold Trans.	believes(player, formula(exists(and(key(K), iscolor(K, red)), inside(K, box2))))
Translation	(10%) believes(player, contains(red, box2))
	(10%) type about(player, box2, contains)
Error	Syntax errors, out of vocabulary term "contains".
Gemini Flash 8B Unconstrained + Majority Vote	
Statement	The player initially did not know if box 2 had a key.
Gold Trans.	not_knows_if(player, formula(exists(key(K), inside(K, box2))))
Translation	(100%) not_knows_if(player, exists(key(K), inside(K, box2)))
Error	Missing "formula" predicate to denote base formula.

Table C2: Erroneous ELoT translations for different parsing methods.

constrained SMC pass the approximate equivalence check, and SMC often produces a distribution over reasonable alternative translations (Row 1). In contrast, unconstrained sampling from the same LLaMa 3.18B base model leads to many syntax errors that majority voting cannot correct for (Row 2). With Gemini Flash 8B, we observe the mode collapse issue that frequently afflicts finetuned LLMs (O'Mahony et al., 2024), often leading all n_σ=10 samples to produce the same incorrect translation (Row 3).

C.2 Unimodal LLM Baseline with Symbolic PDDL Inputs

In our experiment, we used raw visual inputs, such as images or videos, for the multimodal LLM baselines. As the LaBToM model operates over symbolic representations of agent and environment states encoded in the Planning Domain Definition Language (PDDL), we also experimented with prompting LLMs with

Prompt Format	Current Beliefs	Initial Beliefs
Image + Narrative + FS	0.52 (0.01)	0.41 (0.01)
PDDL + Narrative + FS	0.32 (0.01)	0.20 (0.01)

Table C3: GPT-4o correlation with human ratings across different input representations.

symbolic PDDL inputs. We report the results on the best performing LLM baseline (GPT4o I+Na+FS) in Table C3, which shows that replacing raw visual input with PDDL (as text) hinders performance.

C.3 Impact of a Normalized Statement Prior

In the main text, we report results under the assumption that human observers respond as if they have a normalized 50-50 prior U_φ over whether each statement φ is true. Under this assumption, the posterior truth-value of a statement $P(\llbracket\varphi\rrbracket_{(s_t,b_t)}|a_{1:T}, o_{0:T})$ can be interpreted as a *normalized likelihood*:

$$\bar{L}(\llbracket\varphi\rrbracket_{(s_t,b_t)}|a_{1:T}, o_{0:T}) = \frac{P(a_{1:T}, o_{0:T}|\llbracket\varphi\rrbracket_{(s_t,b_t)})}{P(a_{1:T}, o_{0:T}|\llbracket\varphi\rrbracket_{(s_t,b_t)}) + P(a_{1:T}, o_{0:T}|\llbracket\neg\varphi\rrbracket_{(s_t,b_t)})} \quad (14)$$

An alternative assumption is to use a uniform prior $U_{\mathcal{S}_0 \times \mathcal{B}_0}$ over all possible initial states $\mathcal{S}_0$ and belief distributions $\mathcal{B}_0$. This has the effect of up-weighting statements which are true in *more possible worlds*, e.g. *"The player believes that a red key might be in box 1, 2, or 3."*.

Table C4 shows the impact of making either assumption, in terms of the Pearson's correlation coefficient (PCC) r and mean absolute error (MAE) with human judgments, assuming the model has the gold translations. Consistent with Ying et al. (2024b), using a normalized statement prior largely improves the correlation while reducing the mean absolute error for the full LaBToM model. In other words, people appear more willing to say that a statement φ is true only if they have *evidence* for φ, and otherwise default to a 50-50 rating. We see sharper differences for initial belief statements, which is

Model	Prior	Overall		Current		Initial	
		PCC r ↑	MAE ↓	PCC r ↑	MAE ↓	PCC r ↑	MAE ↓
LaBToM (ours)	U_φ	**0.81 (0.01)**	**0.195 (0.004)**	0.81 (0.01)	**0.193 (0.006)**	**0.80 (0.01)**	**0.196 (0.006)**
	$U_{\mathcal{S}_0 \times \mathcal{B}_0}$	0.78 (0.01)	0.226 (0.004)	**0.82 (0.01)**	0.211 (0.006)	0.71 (0.01)	0.244 (0.005)
Non-Planning	U_φ	0.47 (0.01)	0.200 (0.003)	0.60 (0.02)	0.185 (0.004)	0.20 (0.01)	0.219 (0.006)
	$U_{\mathcal{S}_0 \times \mathcal{B}_0}$	0.55 (0.01)	0.252 (0.003)	0.70 (0.01)	0.215 (0.005)	0.34 (0.02)	0.300 (0.005)
True Belief	U_φ	0.12 (0.01)	0.475 (0.003)	0.09 (0.01)	0.482 (0.004)	0.13 (0.02)	0.466 (0.004)
	$U_{\mathcal{S}_0 \times \mathcal{B}_0}$	0.12 (0.01)	0.475 (0.003)	0.09 (0.01)	0.482 (0.004)	0.13 (0.02)	0.467 (0.004)

Table C4: Similarity of human and model ratings for a normalized (U_φ) vs. unnormalized ($U_{\mathcal{S}_0 \times \mathcal{B}_0}$) prior.

likely because priors have a stronger effect on initial beliefs, whereas an agent's current beliefs are more strongly determined by their percepts: If an agent sees that a box is empty, an observer's judgment about whether the agent believes that the box is empty should not depend on the observer's prior.

C.4 Per-Factor Model Performance using Gold Translations

Table C5 shows the correlation between human judgments and model outputs when using the gold ELoT translations. Performance is broken down by the annotated factors described in Table 2. Using gold ELoT translations improves LaBToM's correlation with human judgments from r=0.76 (Table 5) to r=0.81.

Additionally, LaBToM robustly outperforms the baselines on almost all of these data splits, achieving a correlation around r=0.8 in each case. The one exception is the set of statements about what the agent initially knows (*Init. Know.*, $r = 0.31$), such as *"The player initially knew that the red key was in box 3"*. As we discuss in the next section, this is likely because human participants assume that *direct perception* or *justification* is necessary for other agents to know some proposition ϕ. In contrast, our model treats knowledge claims as equivalent to claims of true belief.

Model		Human Correlation r (s.e.)				
		Current	*Curr. Poss.*	*Curr. Prob.*	*Curr. Comp.*	*Curr. Know.*
LaBToM	Full (ours)	**0.81 (0.01)**	**0.80 (0.02)**	**0.76 (0.03)**	**0.76 (0.03)**	**0.89 (0.01)**
	Non-Planning	0.60 (0.02)	0.65 (0.02)	0.48 (0.04)	0.34 (0.04)	0.75 (0.02)
	True Belief	0.09 (0.01)	0.14 (0.03)	0.17 (0.03)	0.02 (0.03)	0.43 (0.02)
GPT-4o	I, Na, FS	0.59 (0.01)	0.62 (0.03)	0.63 (0.03)	0.64 (0.02)	0.67 (0.02)
	I, Na	0.52 (0.01)	0.50 (0.03)	0.54 (0.04)	0.61 (0.03)	0.58 (0.02)
	I, Pl	0.32 (0.01)	0.50 (0.03)	0.30 (0.03)	0.42 (0.02)	0.21 (0.02)
Gemini 1.5 Pro	V, Na, FS	0.28 (0.01)	0.23 (0.03)	0.32 (0.03)	0.10 (0.03)	0.28 (0.02)
	I, Na, FS	0.29 (0.01)	0.25 (0.03)	0.39 (0.03)	0.17 (0.03)	0.26 (0.02)
		Initial	*Init. Poss.*	*Init. Prob.*	*Init. Comp.*	*Init. Know.*
LaBToM	Full (ours)	**0.80 (0.01)**	**0.84 (0.02)**	**0.81 (0.03)**	**0.78 (0.02)**	0.31 (0.05)
	Non-Planning	0.20 (0.01)	0.32 (0.02)	-0.05 (0.03)	0.21 (0.03)	0.34 (0.05)
	True Belief	0.13 (0.02)	-0.09 (0.03)	0.12 (0.05)	-0.03 (0.04)	0.31 (0.05)
GPT-4o	I, Na, FS	0.41 (0.01)	0.41 (0.03)	0.01 (0.03)	0.26 (0.03)	0.31 (0.05)
	I, Na	0.41 (0.01)	0.30 (0.03)	0.48 (0.04)	0.31 (0.02)	**0.51 (0.05)**
	I, Pl	0.18 (0.01)	0.05 (0.03)	0.15 (0.04)	0.23 (0.03)	0.30 (0.05)
Gemini 1.5 Pro	V, Na, FS	0.14 (0.02)	-0.02 (0.03)	-0.04 (0.04)	-0.08 (0.03)	0.27 (0.05)
	I, Na, FS	0.11 (0.02)	0.02 (0.03)	-0.04 (0.04)	-0.02 (0.03)	0.29 (0.05)

Table C5: Human vs. model correlations broken down by linguistic factors. LaBToM results use gold translations.
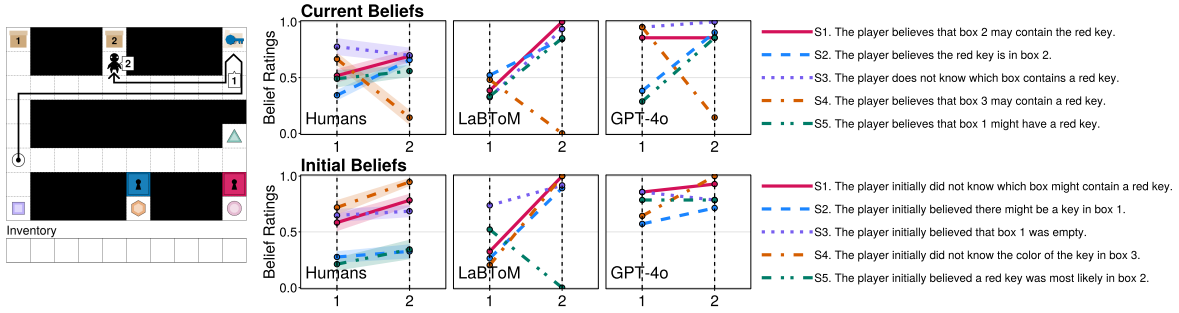
Figure C2: Scenario illustrating differences in human and model ratings of knowledge claims.

C.5 Differences in Human Ratings vs. Model Inferences

While LaBToM largely matches human evaluations of epistemic language both in aggregate and at the individual scenario level, there a number of interesting ways in which they differ.

People are less certain than LaBToM. One difference is simply that our model tends to be more certain than people, using the extremes of the 0-1 scale in ways that our participants tended to avoid. This effect did not appear to be driven by the choice of the Boltzmann inverse temperature β, since lower values of β (which increase model uncertainty) led to poorer fits with human data. Instead, humans may be evaluating the truth a statement ϕ less strictly than our model does, perhaps by maintaining uncertainty over the probability thresholds θ associated with each statement.

People appear to adapt probability thresholds. Threshold uncertainty is closely related to another potential driver of difference: Unlike our model, participants appear to contextually adapt probability thresholds associated with modal words, in line with work on the pragmatics of epistemic modals (Schuster and Degen, 2020; Rudin, 2016; Lassiter, 2017). This is evinced by human responses for current belief statement S2 in Fig. 3b. People rate *"The player is unsure which box has a key"* highly despite the apparent confidence that the player exhibits in looking for a key in box 3 (at the expense of looking in box 1 or box 2). This is consistent with an upwards adjustment of $\theta_{\text{uncertain}}$ from 0.55, such that the player is judged as uncertain even when they seem to think it is quite likely for a key to be in box 3. In contrast, our model thinks it is unlikely that the player is uncertain. Similar effects can be seen for current statement S1 and S5 in the same scenario, except that people appear to adjust their threshold for *may* and *might* downwards from 0.3 and 0.2 to accommodate even highly unlikely possibilities.

People respond as if knowledge requires justification. Unlike our model, which reduces knowledge statements to true belief claims (Table 1), people appear to assume that some form of justification (e.g. via direct perception) is necessary for agents to *know* that some proposition ϕ is true. This is clearly illustrated in Figure C2: In this scenario, the player very confidently walks past boxes 1 and 2 towards box 3, suggesting a strong belief about its contents. Since our model just treats knowledge as true belief (and it is quite possible that the player is correct), LaBToM assigns less than 50% probability to *"The player does not know which box contains a red key"* and *"The player initially did not know the color of the key in box 3"* at Judgment Point 1 (which occurs *before* box 3 is opened). Human judgments differ significantly, assigning more than 50% to these ignorance claims. This is consistent with people understanding knowledge to require justification or direct perception: Since the player has not seen what is in box 3 by Judgment Point 1, they cannot *know* what is in that box.

People reason boundedly about logical implications. A final difference between humans and our model is that people appear to exhibit bounded reasoning when evaluating statements with non-obvious implications. In Figure C2, for example, *"The player initially believed a red key was most likely in box 2"* is rated lowly by humans initially, then more highly, whereas our model assigns a zero rating to that statement by Judgment Point 2. This is because by then, it is clear that the player did not need a blue key, and was instead looking for a red key in box 1. The fact that they looked in box 1 first also implies that they initially thought box 1 was most likely to contain a red key, not box 2. Our model captures this reasoning, but people do not seem to independently grasp these multi-step implications, consistent with studies on the boundedness of human reasoning (Smets and Solaki, 2018; Mercier and Sperber, 2017).