AutoToM: Automated Bayesian Inverse Planning and Model Discovery for Open-ended Theory of Mind

Anonymous ACL submission

Abstract

Theory of Mind (ToM), the ability to understand people's mental variables based on their behavior, is key to developing socially intelligent agents. Current approaches to Theory of Mind reasoning either rely on prompting Large Language Models (LLMs), which are prone to systematic errors, or use rigid, handcrafted Bayesian Theory of Mind (BToM) models, which are more robust but cannot generalize across different domains. In this work, we introduce *AutoToM*, an automated Bayesian Theory of Mind method for achieving open-ended machine Theory of Mind. *AutoToM* can operate in any domain, infer any mental variable, and conduct robust Theory of Mind reasoning of any order. Given a Theory of Mind inference problem, *AutoToM* first proposes an initial BToM model. It then conducts automated Bayesian inverse planning based on the proposed model, leveraging an LLM as the backend. Based on the uncertainty of the inference, it iteratively refines the model, by introducing additional mental variables and/or incorporating more timesteps in the context. Empirical evaluations across multiple Theory of Mind benchmarks demonstrate that *AutoToM* consistently achieves state-of-the-art performance, offering a scalable, robust, and interpretable approach to machine Theory of Mind.

1 Introduction

To successfully engage in rich and complex social interactions such as cooperation, communication, and social learning, humans must adequately understand one another's mental states (e.g., goals, beliefs, desires). This ability is termed Theory of Mind (ToM) (Wimmer and Perner, 1983). Prior works have demonstrated that like human interactions, Theory of Mind is also crucial for the success of human-AI interactions (e.g., Dautenhahn, 2007; Hadfield-Menell et al., 2016; Liu et al., 2018). In particular, to safely and productively interact with humans in an open-ended manner, AI systems need to interpret humans' mental states from observed human behavior (e.g., Chandra et al., 2020; Wang et al., 2021; Wan et al., 2022; Patel and Chernova, 2022; Puig et al., 2023; Zhi-Xuan et al., 2024; Ying et al., 2024).

There are two primary approaches to developing machine Theory of Mind in recent works. First, with the rapid progress of large language models (LLMs), there has been an increasing interest in directly applying LLMs to reason about people's mental states with prompting strategies such as perspective-taking (Wilf et al., 2023; Sclar et al., 2023; Jung et al., 2024), change-tracking (Huang et al., 2024), and temporal-spatial reasoning (Hou et al., 2024). However, even with these advanced prompting techniques, state-of-the-art LLMs still make systematic errors in complex scenarios (Jin et al., 2024). Second, cognitive studies have demonstrated that model-based inference, in particular, Bayesian inverse planning (BIP), can reverse engineer human-like theory of Mind reasoning (Baker et al., 2009; Ullman et al., 2009; Baker et al., 2017; Zhi-Xuan et al., 2020). BIP relies on Bayesian Theory of Mind (BToM) models (Baker et al., 2017) to approximate rational agent behaviors. Inspired by this, recent works have proposed to combine BIP and LLMs to achieve scalable yet robust model-based ToM inference (Jin et al., 2024; Shi et al., 2024). While these methods significantly outperform LLMs in specific domains, they typically require manual specification of BToM models, including necessary mental variables (e.g., goals, beliefs) for answering a given ToM question. Therefore, they lack the required generalizability for open-ended Theory of Mind.

In this work, we aim to develop a fully *automated* and *open-ended* Theory of Mind reasoning method. That is a unified method that can be applied to robustly infer any given mental variable in any domain. Achieving this aim requires address-

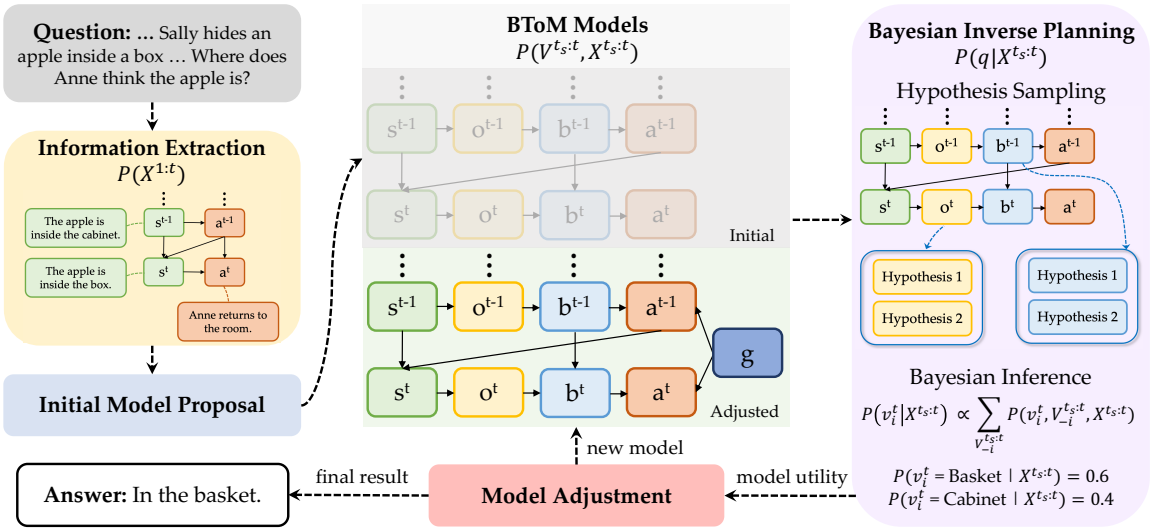

Figure 1: An overview of *AutoToM*. $X^{t_s:t}$ are observable variables, $V^{t_s:t}$ are latent mental variables, and q is the query (in this case, a mental variable $v_i^t \in V^t$). $t_s : t$ denotes timesteps from t_s to t in the context that are considered for inference. Variables s^t, o^t, b^t, a^t, g^t represent state, observation, belief, action, and goal, respectively, with solid arrows indicating dependencies defined in the models. Given a question, we extract the observable variables (information extraction) and propose an initial BToM model. This is followed by automated Bayesian inverse planning and iterative model adjustment. When the model utility is high enough, we will produce the final answer based on the inference result.

ing two critical questions: (1) How can we ensure that our approach is flexible enough to adapt across contexts, robust enough to model diverse human behaviors, and scalable enough to tackle increasingly complex scenarios? (2) How can we avoid manually defining model structures and instead autonomously discover the appropriate model for mental inference?

To address these challenges, we introduce *AutoToM*, a general framework for open-ended Theory of Mind. It automates every aspect of Bayesian inverse planning, including the proposal and adjustment of model structures, the identification of relevant timesteps, the generation of hypotheses, and the execution of Bayesian inference. It is designed to operate in *any context*, infer *any mental state*, reason about *any number of agents*, and support *any order of recursive reasoning*, which represents our vision of an open-ended and robust machine Theory of Mind.

Figure 1 provides an overview of *AutoToM*, which consists of two main components:

First, **Automated Bayesian Inverse Planning.** *AutoToM* is capable of flexibly modeling various mental variables and their dependencies for any specified BToM model (in the form of a Bayesian network). The construction, information flow, and computations within a given BToM model are entirely automated, leveraging the adaptability of the LLM backend. Specifically, conditioned on observable variables and their values extracted from the context (by an LLM), *AutoToM* samples a small set of hypotheses for each latent mental variable using an LLM. Given the hypotheses, *AutoToM* then conducts Bayesian inference to produce the posterior distribution of the target mental variable in the question. To achieve this, *AutoToM* leverages an LLM to estimate each local conditional in the BToM model. (Section 3.3)

Second, **Automated Model Discovery.** In a given scenario, *AutoToM* performs automated model proposals and iteratively adjusts variables and the timesteps of observable variables. We ground the BToM model proposals in cognitive models of human decision-making (e.g., Baker et al., 2017; Ullman et al., 2009). The goal is to include the relevant mental variables and timesteps necessary for the inference, optimizing based on model utility, which balances the certainty of the inference and the complexity of the model. This approach eliminates the need for manual effort in defining model structures and enhances generalization by enabling automatic adaptation to diverse scenarios. Furthermore, *AutoToM* can select a different suitable model for each timestep, enabling it to adapt dynamically to changing circumstances. (Section 3.4)

AutoToM is the first model-based ToM method that extends beyond domain-specific applications and addresses open-ended scenarios. It integrates

the flexibility of LLMs with the robustness of Bayesian inverse planning. We evaluate *AutoToM* in multiple ToM benchmarks. The results consistently show that *AutoToM* achieves state-of-the-art performance, establishing a scalable, robust, and interpretable framework for machine ToM.

2 Related Works

Enhancing LLMs' Theory of Mind. There has been systematic evaluation that revealed LLMs' limitations in achieving robust Theory of Mind inference (Ullman, 2023; Shapira et al., 2023). To enhance LLMs' Theory of Mind capacity, recent works have proposed various prompting techniques. For instance, SimToM (Wilf et al., 2023) encourages LLMs to adopt perspective-taking, PercepToM (Jung et al., 2024) improves perception-to-belief inference by extracting relevant contextual details, and Huang et al. (2024) utilize an LLM as a world model to track environmental changes and refine prompts. Explicit symbolic modules also seem to improve LLM's accuracy through dynamic updates based on inputs. Specifically, TimeToM (Hou et al., 2024) constructs a temporal reasoning framework to support inference, while SymbolicToM (Sclar et al., 2023) uses graphical representations to track characters' beliefs. Additionally, Wagner et al. (2024) investigates ToM's necessity and the level of recursion required for specific tasks. However, these approaches continue to exhibit systematic errors in long contexts, complex behaviors, and recursive reasoning due to inherent limitations in inference and modeling (Jin et al., 2024; Shi et al., 2024). Most of them rely on domain-specific designs, lacking open-endedness.

Model-based Theory of Mind inference. Model-based Theory of Mind inference, in particular, Bayesian inverse planning (BIP) (Baker et al., 2009; Ullman et al., 2009; Baker et al., 2017; Zhi-Xuan et al., 2020), explicitly constructs representations of agents' mental states and how mental states guide agents' behavior via Bayesian Theory of Mind (BToM) models. These methods can reverse engineer human ToM inference in simple domains (e.g., Baker et al., 2017; Netanyahu et al., 2021; Shu et al., 2021). Recent works have proposed to combine BIP with LLMs to achieve robust ToM inference in more realistic settings (Jin et al., 2024; Shi et al., 2024). However, these methods require manual specification of the BToM models as well as rigid, domain-specific implementations of Bayesian inference, limiting their adaptability to open-ended scenarios. To overcome this limitation, we propose *AutoToM*, a method capable of automatically modeling mental variables across diverse conditions and conducting automated BIP without domain-specific knowledge or implementations.

Automated Modeling with LLMs. There has been an increasing interest in integrating LLMs with inductive reasoning and probabilistic inference for automated modeling. Piriyakulkij et al. (2024) combine LLMs with Sequential Monte Carlo to perform probabilistic inference about underlying rules. Iterative hypothesis refinement techniques (Qiu et al., 2023) further enhance LLM-based inductive reasoning by iteratively proposing, selecting, and refining textual hypotheses of rules. Beyond rule-based hypotheses, Wang et al. (2023) prompt LLMs to generate natural language hypotheses that are then implemented as verifiable programs, while Li et al. (2024) propose a method in which LLMs construct, critique, and refine statistical models represented as probabilistic programs for data modeling. Cross et al. (2024) leverage LLMs to propose and evaluate agent strategies for multi-agent planning but do not specifically infer individual mental variables. Our method also aims to achieve automated modeling with LLMs. Unlike prior works, we propose a novel automated model discovery approach for Bayesian inverse planning, where the objective is to confidently infer any mental variable given any context via constructing a suitable Bayesian Theory of Mind model.

3 AutoToM

3.1 Preliminaries

Bayesian Inverse Planning (BIP) is a computational framework that models how observers infer unobservable mental states—such as beliefs and goals—from an agent's behavior (Baker et al., 2009). It assumes that the agent acts rationally according to a generative model, a Bayesian Theory of Mind (BToM) model (Baker et al., 2017), which specifies how internal variables lead to observable actions in a Bayesian network (e.g., the example models on the bottom panels in Figure 2). Using inverse inference, BIP inverts this generative process to assess what latent mental variables can lead to observed agent behavior. This probabilistic inference reasons about how agents make decisions, serving as a robust solution to ToM challenges.

There have been different instantiations of BIP

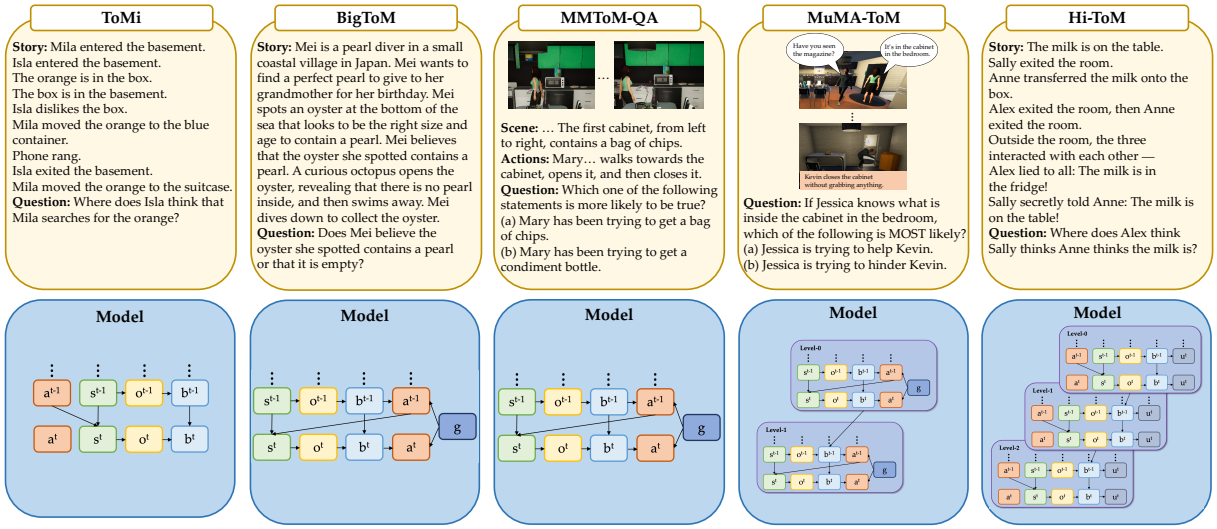

Figure 2: Examples questions (top panels) and the necessary Bayesian Theory of Mind (BToM) model for Bayesian inverse planning (bottom panels) in diverse Theory of Mind benchmarks. *AutoToM* aims to answer any Theory of Mind question in a variety of benchmarks, encompassing different mental variables, observable contexts, numbers of agents, the presence or absence of utterances, wording styles, and modalities. It proposes and iteratively adjusts an appropriate BToM and conducts automated Bayesian inverse planning based on the model. There can be more types of questions/models in each benchmark beyond the examples shown in this figure.

in prior works (e.g., Baker et al., 2009; Ullman et al., 2009; Ong et al., 2019; Jha et al., 2024). Here we formally define BIP in a unified way. We denote the observable variables at time t describing the environment and an agent's behaviors as $X^t = \{x_i^t\}_{i \in N_X}$, where N_X is the set of observable variables and x_i^t is a particular variable (state, action, or utterance) at t. We can extract the values of these observable variables from the context provided in a ToM problem. We denote an agent's latent mental variables at time t as $V^t = \{v_i^t\}_{i \in N_V}$, where N_V is the set of mental variables and v_i^t is a particular mental variable (e.g., goal, desire, belief) at t. BIP formulates a BToM model as a Bayesian network that defines $P(V^t, X^t)$, which indicates how the mental variables drive an agent's behavior. Given this model, BIP infers the latent mental variables for the current step t:

$$P(V^t|X^t) = \frac{P(V^t, X^t)}{\sum_V P(V, X^t)} \propto P(V^t, X^t). \quad (1)$$

In many real-world scenarios, past observations (such as actions taken at the previous steps) are often valuable for inferring the mental variables at the current step. Suppose the context from step t_s to step t is relevant for the current mental variable inference, then the inference becomes:

$$P(V^{t_s:t}|X^{t_s:t}) \propto P(V^{t_s:t}, X^{t_s:t}). \quad (2)$$

In a ToM problem, there is a query concerning a specific target variable q to be inferred. We can answer the query via $P(q|X^{t_s:t})$. Typically, the query asks about a latent mental variable $q = v_i^t \in V^t$, the posterior probability is obtained by marginalizing over other latent variables $V_{-i}^{t_s:t}$ which is the subset of $V^{t_s:t}$ excluding v_i^t:

$$P(v_i^t|X^{t_s:t}) \propto \sum_{V_{-i}^{t_s:t}} P(v_i^t, V_{-i}^{t_s:t}, X^{t_s:t}). \quad (3)$$

This can also be extended to predicting a future observable variable $q = x_i^{t+1}$ given observations from t_s to t:

$$P(x_i^{t+1}|X^{t_s:t}) \propto \sum_{V^{t_s:t}} P(V^{t_s:t}, x_i^{t+1}, X^{t_s:t}). \quad (4)$$

To conduct BIP in different scenarios, we must formulate the mental variables and their causal relationships with agent behavior using suitable BToM models. Each model M is uniquely defined by the observable variables and the latent mental variables, i.e., $M = (V^{t_s:t}, X^{t_s:t})$. Let $s^t \in S$ be the state at time t, and $a^t \in A$ be the action taken by the agent at time t. The current state and action determines the next state s^{t+1}. When the agent has an explicit goal $g \in G$, this setup constitutes a Markov Decision Process (MDP). If the agent only has a partial observation of the state, the model becomes a Partially Observable Markov Decision Process (POMDP) (Kaelbling et al., 1998). In POMDP, the agent receives a partial observation o^t of the true state s^t, maintains a belief b^t over the possible states, and selects its action a^t based on this

belief and goal. When there is high-order recursive reasoning between two agents (i and j), we can adopt an Interactive POMDP (I-POMDP) (Gmytrasiewicz and Doshi, 2005), where the belief of state at level $l > 0$ for agent i will become the belief of interactive state $is^t = (s, b_{j,l-1}, g_j)$, where $b_{j,l-1}$ is the belief of agent j at the lower level $l-1$ and g_j is agent j's goal.

For instance, given a POMDP model, we can conduct the following Bayesian inference to infer the agent's belief b^t at time t from the observed state s^t and a^t:

$$P(b^t \mid s^t, a^t) \propto \sum_{b^{t-1}} \sum_{o^t} \sum_{g} P(a^t \mid b^t, g) \cdot P(b^t \mid b^{t-1}, o^t) P(o^t \mid s^t) \cdot P(b^{t-1}) P(g). \quad (5)$$

3.2 Overview of *AutoToM*

As shown in Figure 1, *AutoToM* aim to construct a suitable BToM model for Bayesian inverse planning to confidently infer any target variable. There are several key challenges in achieving this: First, different ToM inference problems require different BToM models (as illustrated in Figure 2); our model does not know which is most suitable a priori. Second, in a given context, our method must determine which time steps are relevant. Third, there is no predefined hypothesis space for each mental variable, and each space could be infinite. Last, to infer mental variables in any context, our method must flexibly represent them without assuming specific types of representations.

AutoToM addresses these challenges in the two key components: (1) automated Bayesian inverse planning which conducted Bayesian inverse planning given a specified BToM model and (2) automated model discovery which proposes and adjusts the BToM model based on the question and the inference results. These two components form a self-improvement loop to iteratively update the BToM model and corresponding inference result as summarized in Algorithm 1. We discuss these two components in Section 3.3 and Section 3.4 respectively. More details are provided in Appendix B.

3.3 Automated Bayesian Inverse Planning

Given a BToM model, M, including the necessary latent mental variables $V^{t_s:t}$ and the observable variables $X^{t_s:t}$, we integrate LLMs as the computational backend to implement every aspect of the Bayesian inverse planning (Line 2-6 in Algorithm 1). In particular, the hypothesis sampling module suggests a small set of possible values of latent variables. The Bayesian inference module then computes the posterior distribution of the target variable in the query based on Eqn. (3) or Eqn.(4).

Hypothesis Sampling. Conventional BIP assumes a manually defined hypothesis space and hypothesis representation for each latent mental variable. Our hypothesis sampling module instead leverages an LLM to propose only a small set of quality hypotheses for each latent variable in $V^{t_s:t}$. This is similar to amortized inference (Ritchie et al., 2016; Jha et al., 2024) but does not require learning a data-driven proposal distribution. To ensure that the sampled hypotheses are relevant to the ToM inference problem, we guide the sampling process with both the question and the observable variables $X^{t_s:t}$. To remove spurious hypotheses generated by the LLM, we further apply *hypothesis reduction* to eliminate unlikely hypotheses and reduce the hypothesis space. Unlikely hypotheses are identified by evaluating the local conditionals. For instance,

Algorithm 1 *AutoToM*

Require: Question Q, terminate threshold U_{min}
1: ▷ Automated Bayesian inverse planning
2: **function** BIP($M = (V^{t_s:t}, X^{t_s:t}), q$)
3: **Sample** hypotheses for latent variables $V^{t_s:t}$
4: **Conduct** Bayesian inference via LLMs to compute $P(q \mid^{t_s:t})$ ▷ Based on Eqn. (3) or Eqn. (4)
5: **return** $P(q \mid X^{t_s:t})$
6: **end function**
7: ▷ Automated Model Discovery
8: **Extract** query q from Q
9: **Extract** observable variables $X^{1:t}$ from Q
10: $t_s \leftarrow t$
11: **while** $t_s \geq 1$ **do**
12: **Propose** initial V^{t_s}
13: $M \leftarrow (V^{t_s:t}, X^{t_s:t})$
14: $P(q \mid X^{t_s:t}) \leftarrow \text{BIP}(M, q)$
15: **Compute** the model utility $U(M, q)$
16: **while** V^{t_s} does not contain all mental variables **do**
17: $v^{t_s}_{\text{new}} = \arg\max_{v \notin V^{t_s}} U(M + v, q)$ ▷ Based on results from BIP($M + v, q$)
18: **if** $U(M + v^{t_s}_{\text{new}}, q) > U(M, q)$ **then**
19: $M \leftarrow M + v^{t_s}_{\text{new}}$
20: $P(q \mid X^{t_s:t}) \leftarrow \text{BIP}(M, q)$
21: **else**
22: **Exit** loop
23: **end if**
24: **end while**
25: **if** $U(M, q) \geq U_{\text{min}}$ **then**
26: **Exit** loop
27: **else**
28: $t_s \leftarrow t_s - 1$
29: **end if**
30: **end while**
31: **Return** the answer $A \leftarrow \arg\max_q P(q \mid X^{t_s:t})$

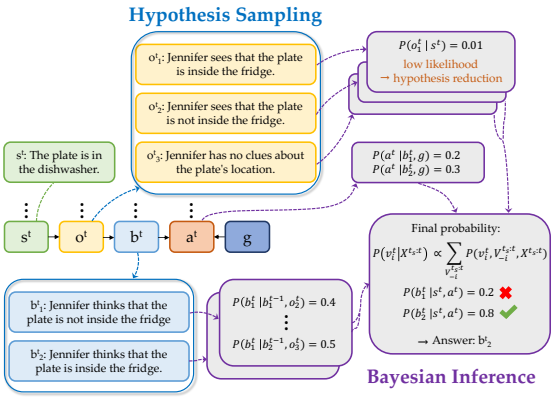

Figure 3: Illustration of automated Bayesian inverse planning given a BToM model. We sample hypotheses for each latent variable (o^t and b^t in this example), remove spurious hypotheses, and finally conduct Bayesian inference based on estimated local conditionals.

we discard observation hypotheses with low likelihood conditioned on the state as shown in Figure 3.

Bayesian Inference. As shown in Figure 3, we estimate each local conditional in $P(V^{t_s:t}, X^{t_s:t})$ using an LLM. After marginalizing the joint distribution over non-target latent variables, we then produce the posterior probabilities of the target variable, i.e., Eqn. (3). This also applies to predicting a future observable variable, i.e., Eqn. (4).

Our automated Bayesian inverse planning greatly generalizes prior methods that combine BIP and LLMs, such as BIP-ALM (Jin et al., 2024) and LIMP (Shi et al., 2024). Specifically, prior methods assume a fixed model structure for a few specific ToM inference problems. They also cannot propose hypotheses for non-target latent variables. In contrast, *AutoToM* can conduct any ToM inference based on any BToM model structure and consider multiple non-target latent variables simultaneously. Additionally, unlike prior methods, our Bayesian inference can work with arbitrary levels of recursive for high-order ToM inference.

3.4 Automated Model Discovery

Prior works on Bayesian inverse planning rely on manually designed BToM models, which limits their applicability to domain-specific scenarios. In contrast, the Automated Model Discovery component automatically proposes a model and dynamically adjusts it to ensure both the *effectiveness* of the model—confidently inferring agents' mental states—and the *efficiency* of the inference by minimizing model complexity. To achieve this, we formulate the utility of a model $M = (V^{t_s:t}, X^{t_s:t})$ used for answering a given query q as

$$U(M, q) = R(M, q) - C(M), \tag{6}$$

where $R(M, q)$ assesses the model's confidence in answering the query, and $C(M)$ is its computational cost. In this work, the reward is defined as $R(M, q) = -H(P(q|X^{t_s:t}))$, where $P(q|X^{t_s:t})$ is the probability distribution of the target variable based on Eqn. (3) or Eqn. (4), and $H(\cdot)$ is its entropy. This is designed to decrease the uncertainty in the inference. To minimize the compute needed for the inference, we define the cost of the model as $C(M) = \alpha|M|$, where $|M|$ denotes the model's complexity, measured by the number of latent mental variables, and $\alpha > 0$ is a weighting factor. The cost increases with complexity, encouraging parsimonious models with lower compute.

There are three modules for Automated Model Discovery:

Information Extraction. The information extraction module (Line 9 in Algorithm 1) processes the context to identify the values of observable variables $X^{1:t}$, including states (s^t), actions (a^t), and utterances (u^t), organized along a timeline (the number of timesteps is determined by the number of actions and utterances). When there are multiple agents, we identify whose mental state the question is asking about (i.e., the target agent), and then construct the timesteps based on the target agent's actions and/or utterances. The extraction is performed once using an LLM and used for model proposal and Bayesian inverse planning.

Initial Model Proposal. We employ an LLM to propose an initial BToM model based on $X^{1:t}$ and the query (Line 12-15 in Algorithm 1). This initial model represents a minimal model, containing only the essential mental variables needed to answer the question. This initial proposal also includes assessing the level of recursive reasoning necessary for higher-order ToM inference. Note that we always begin with only considering the last timestep in context, i.e., $t_s = t$. Following this model, we conduct automated Bayesian inverse planning, as described in Section 3.3. If the model utility exceeds a threshold U_{min}, we accept the inference result as the final answer. Otherwise, we use the model utility to guide model adjustments.

Model Adjustment. We iteratively adjust the proposed model to maximize the utility (Line 11-30 in Algorithm 1) by considering two types of model adjustments: variable adjustment (Figure 4**A**) and timestep adjustment (Figure 4**B**):

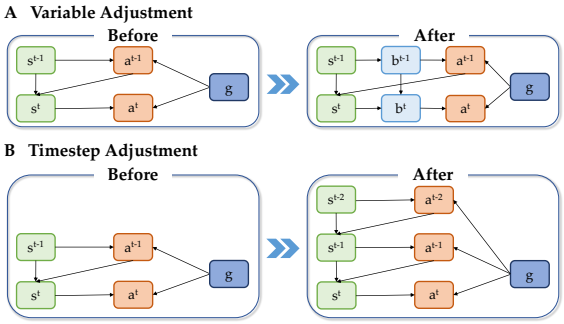

Figure 4: Given any ToM inference problem, we automatically refine the BToM model by alternating between (**A**) variable adjustment (introducing belief in this example) and (**B**) timestep adjustment.

Variable Adjustment. We refine the model structure at a specific timestep by iteratively introducing new, relevant latent variables into the model to address uncertainty in the inference. These variables include goal, belief, observation, and interactive state as summarized in Table 10 in Appendix B. This follows the typical causal structures introduced in prior decision-making models (e.g., Kaelbling et al., 1998; Baker et al., 2017; Ullman et al., 2009; Gmytrasiewicz and Doshi, 2005). Such restricted variable adjustment helps reduce the model space and ensures the proposed models can explain human behavior. For each adjustment, we compute the updated model utility and accept the modification that offers the biggest increase in utility. This iterative process continues until no further significant improvements are possible. Note that our method can still propose diverse models beyond standard MDP, POMDP, and I-POMDP even with this restricted model adjustment. Appendix B.5 provides more details on the model space.

Timestep Adjustment. If model utility remains low and no significant improvement can be achieved via variable adjustment within the current timesteps $t_s : t$, we incorporate an additional step, $t_s - 1$, to enhance context for inference. Upon adding a timestep, we first apply the initial model structure and then adjust variables accordingly.

We iterate the variable and timestep adjustments, as outlined in Algorithm 1, until either the model utility exceeds the desired threshold or no further meaningful improvement is possible.

4 Experiments

4.1 Experimental Settings

We evaluated our method on multiple Theory of Mind benchmarks, including ToMi (Le et al., 2019), BigToM (Gandhi et al., 2024), MMToM-QA (Jin et al., 2024), MuMA-ToM (Shi et al., 2024), and Hi-ToM (He et al., 2023). The diversity and complexity of these benchmarks pose significant reasoning challenges. For instance, MMToM-QA and MuMA-ToM incorporate both visual and textual input, while MuMA-ToM and Hi-ToM require higher-order inference. Additionally, MMToM-QA features exceptionally long contexts, and BigToM presents open-ended scenarios.

Besides the full *AutoToM* method, we additionally evaluated *AutoToM* given manually specified models (AutoToM w/ Model Spec.).

We compared *AutoToM* against state-of-the-art baselines:

LLMs: Llama 3.1 70B (Dubey et al., 2024), Gemini 2.0 Flash, Gemini 2.0 Pro (Team et al., 2023) and GPT-4o (Achiam et al., 2023);

ToM prompting for LLMs: SymbolicToM (Sclar et al., 2023), SimToM (Wilf et al., 2023), TimeToM (Hou et al., 2024), and PercepToM (Jung et al., 2024);

Model-based inference: BIP-ALM (Jin et al., 2024) and LIMP (Shi et al., 2024).

For multimodal benchmarks, MMToM-QA and MuMA-ToM, we adopt the information fusion methods proposed by Jin et al. (2024) and Shi et al. (2024) to fuse information from visual and text inputs respectively. The fused information is in text form. We ensure that all methods use the same fused information as their input.

We use GPT-4o as the LLM backend for *AutoToM* and all ToM prompting and model-based inference baselines to ensure a fair comparison—except for TimeToM, which relies on GPT-4 and is not open-sourced.

4.2 Results

The main results are summarized in Table 1. Unlike *AutoToM*, many recent ToM baselines can only be applied to specific benchmarks. Among general methods, *AutoToM* achieves state-of-the-art results across all benchmarks. In particular, it outperforms its LLM backend, GPT-4o, by a large margin. This is because Bayesian inverse planning is more robust for inferring mental states given long contexts with complex environments and agent behavior. It is also more adept at recursive reasoning which is key to higher-order inference. Notably, *AutoToM* performs comparably to manually specified models, showing that automatic model discovery without domain knowledge is as effective as human-provided models. We provide additional

Method	Type	ToMi	BigToM	MMToM-QA	MuMA-ToM	Hi-ToM	All
SymbolicToM	Specific	**98.60**	-	-	-	-	-
TimeToM	Specific	87.80	-	-	-	-	-
PercepToM	Specific	82.90	-	-	-	-	-
BIP-ALM	Specific	-	-	76.70	33.90	-	-
LIMP	Specific	-	-	-	76.60	-	-
AutoToM w/ Model Spec.	Specific	88.80	**86.75**	**79.83**	**84.00**	**74.00**	**82.68**
Llama 3.1 70B	General	72.00	77.83	43.83	55.78	35.00	47.41
Gemini 2.0 Flash	General	66.70	82.00	48.00	55.33	52.50	60.91
Gemini 2.0 Pro	General	71.90	86.33	50.84	62.22	57.50	65.76
GPT-4o	General	77.00	82.42	44.00	63.55	50.00	63.39
SimToM	General	79.90	77.50	51.00	47.63	71.00	65.41
AutoToM	General	**88.30**	**86.92**	**75.50**	**81.44**	**72.50**	**80.93**

Table 1: Results of *AutoToM* and baselines on all benchmarks. There are two groups of methods: methods that require domain-specific knowledge (e.g., AutoToM w/ Model Spec.) or implementations (e.g., SymbolicToM) and methods that can be generally applied to any domain. "-" indicates that the domain-specific method is not applicable to the benchmark. The best results for each method type are highlighted in bold.

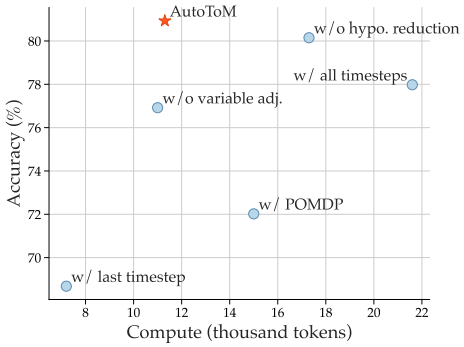

Figure 5: Averaged performance and compute of the full *AutoToM* method (star) and the ablated methods (circles) on all benchmarks.

results and qualitative examples in Appendix A.

4.3 Ablated Study

We evaluated the following variants of *AutoToM* for an ablation study: no hypothesis reduction (**w/o hypo. reduction**); always using POMDP (**w/ POMDP**); always using the initial model proposal without variable adjustment (**w/o variable adj.**); only considering the last timestep (**w/ last timestep**); and considering all timesteps without timestep adjustment (**w/ all timesteps**).

The results in Figure 5 show that the full *AutoToM* method constructs a suitable BToM model, enabling rich ToM inferences while reducing compute. We analyze key model components below:

Hypothesis reduction. Compared to the full method, *AutoToM* w/o hypo. reduction has a similar accuracy but consumes 53% more tokens on average, demonstrating that hypothesis reduction optimizes efficiency without sacrificing performance.

Variable adjustment. *AutoToM* dynamically identifies relevant variables for ToM inference, generalizing domain-specific BIP approaches to open-ended scenarios. Compared to its variant without variable adjustment, *AutoToM* improves performance with minimal additional compute. The variant that always uses POMDP performs well in scenarios aligned with the POMDP assumption (e.g., MMToM-QA) but generalizes poorly elsewhere and incurs much higher computational costs.

Timestep adjustment. By selecting relevant steps for inference, timestep adjustment enhances performance by focusing on essential information. In contrast, the variant using only the last timestep misses crucial details, significantly lowering performance. The variant incorporating all timesteps suffers from higher computational costs and reduced accuracy due to conditioning on unnecessary, potentially distracting information.

Full ablation results are provided in Appendix A.3.

5 Conclusion

We have proposed *AutoToM*, a novel framework for open-ended Theory of Mind. Given any ToM inference problem, *AutoToM* can automatically construct a suitable BToM model and conduct automated Bayesian inverse planning with an LLM backend. Our experimental results demonstrated that *AutoToM* can answer different Theory of Mind questions in diverse scenarios, significantly outperforming baselines. *AutoToM* suggests a promising direction toward cognitively grounded Theory of Mind modeling that is scalable, robust, and open-ended. In the future, we intend to further improve the robustness of *AutoToM* while reducing its inference cost by exploring the possibility of implicit model proposal and Bayesian inference.

Limitations

AutoToM still makes mistakes in several aspects of the inference and model discovery. First, it sometimes proposes hypotheses unrelated to the ToM inference problem, particularly in questions where the definitions of certain mental variables are more ambiguous. Second, the LLM backend may also produce inaccurate likelihood estimation when there are multiple similar hypotheses for a latent variable. Last, model adjustment may fail to recognize the relevance of certain mental variables, resulting in an insufficient model. In addition, while *AutoToM* can balance accuracy and cost to a certain degree, it still requires multiple API calls. For applications with a strict computational budget, there is a need for further reducing the cost.

Ethics Statement

Engineering machine Theory of Mind is an important step toward building socially intelligent AI systems that can safely and productively interact with humans in the real world. Our work provides a novel framework for achieving open-ended and reliable machine Theory of Mind, which may serve as a component of any AI systems designed to interact with humans. The explicit BToM model discovered by *AutoToM* offers an interpretable explanation of the model results, enabling human users to examine and diagnose the model inference. While we do not foresee any negative impact or risk of our work, we acknowledge the importance of robust and trustworthy machine Theory of Mind. Interpretable and cognitively grounded machine Theory of Mind methods such as *AutoToM* may help mitigate the negative effects of LLMs, including hallucinations and biases. Additionally, current Theory of Mind benchmarks are typically constructed using procedurally generated stories and questions. There is a need to carefully examine the potential biases in these benchmarks, to ensure that the models evaluated on these benchmarks are fair and unbiased.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Akshatha Arodi and Jackie Chi Kit Cheung. 2021. Textual time travel: A temporally informed approach to theory of mind. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4162–4172.

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):0064.

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

Rohan Chandra, Aniket Bera, and Dinesh Manocha. 2020. Stylepredict: Machine theory of mind for human driver behavior from trajectories. *arXiv preprint arXiv:2011.04816*.

Logan Cross, Violet Xiang, Agam Bhatia, Daniel LK Yamins, and Nick Haber. 2024. Hypothetical minds: Scaffolding theory of mind for multi-agent tasks with large language models. *arXiv preprint arXiv:2407.07086*.

Kerstin Dautenhahn. 2007. Socially intelligent robots: dimensions of human–robot interaction. *Philosophical transactions of the royal society B: Biological sciences*, 362(1480):679–704.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. 2024. Understanding social reasoning in language models with language models. *Advances in Neural Information Processing Systems*, 36.

Piotr J Gmytrasiewicz and Prashant Doshi. 2005. A framework for sequential planning in multi-agent settings. *Journal of Artificial Intelligence Research*, 24:49–79.

Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. 2016. Cooperative inverse reinforcement learning. In *Advances in neural information processing systems*.

Yinghui He, Yufan Wu, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. 2023. Hi-tom: A benchmark for evaluating higher-order theory of mind reasoning in large language models. *arXiv preprint arXiv:2310.16755*.

Guiyang Hou, Wenqi Zhang, Yongliang Shen, Linjuan Wu, and Weiming Lu. 2024. Timetom: Temporal space is the key to unlocking the door of large language models' theory-of-mind. *arXiv preprint arXiv:2407.01455*.

X Angelo Huang, Emanuele La Malfa, Samuele Marro, Andrea Asperti, Anthony Cohn, and Michael Wooldridge. 2024. A notion of complexity for theory

of mind via discrete world models. *arXiv preprint arXiv:2406.11911*.

Kunal Jha, Tuan Anh Le, Chuanyang Jin, Yen-Ling Kuo, Joshua B Tenenbaum, and Tianmin Shu. 2024. Neural amortized inference for nested multi-agent reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

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

Chani Jung, Dongkwan Kim, Jiho Jin, Jiseon Kim, Yeon Seonwoo, Yejin Choi, Alice Oh, and Hyunwoo Kim. 2024. Perceptions to beliefs: Exploring precursory inferences for theory of mind in large language models. *arXiv preprint arXiv:2407.06004*.

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.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. 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)*, pages 5872–5877.

Michael Y Li, Emily B Fox, and Noah D Goodman. 2024. Automated statistical model discovery with language models. *arXiv preprint arXiv:2402.17879*.

Chang Liu, Jessica B Hamrick, Jaime F Fisac, Anca D Dragan, J Karl Hedrick, S Shankar Sastry, and Thomas L Griffiths. 2018. Goal inference improves objective and perceived performance in human-robot collaboration. *arXiv preprint arXiv:1802.01780*.

Aviv Netanyahu, Tianmin Shu, Boris Katz, Andrei Barbu, and Joshua B Tenenbaum. 2021. Phase: Physically-grounded abstract social events for machine social perception. In *Proceedings of the aaai conference on artificial intelligence*, volume 35, pages 845–853.

Desmond C Ong, Jamil Zaki, and Noah D Goodman. 2019. Computational models of emotion inference in theory of mind: A review and roadmap. *Topics in cognitive science*, 11(2):338–357.

Maithili Patel and Sonia Chernova. 2022. Proactive robot assistance via spatio-temporal object modeling. *arXiv preprint arXiv:2211.15501*.

Wasu Top Piriyakulkij, Cassidy Langenfeld, Tuan Anh Le, and Kevin Ellis. 2024. Doing experiments and revising rules with natural language and probabilistic reasoning. *arXiv preprint arXiv:2402.06025*.

Xavier Puig, Tianmin Shu, Joshua B Tenenbaum, and Antonio Torralba. 2023. Nopa: Neurally-guided online probabilistic assistance for building socially intelligent home assistants. *arXiv preprint arXiv:2301.05223*.

Linlu Qiu, Liwei Jiang, Ximing Lu, Melanie Sclar, Valentina Pyatkin, Chandra Bhagavatula, Bailin Wang, Yoon Kim, Yejin Choi, Nouha Dziri, and 1 others. 2023. Phenomenal yet puzzling: Testing inductive reasoning capabilities of language models with hypothesis refinement. *arXiv preprint arXiv:2310.08559*.

Daniel Ritchie, Paul Horsfall, and Noah D Goodman. 2016. Deep amortized inference for probabilistic programs. *arXiv preprint arXiv:1610.05735*.

Maarten Sap, Ronan LeBras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large lms. *arXiv preprint arXiv:2210.13312*.

Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. 2023. Minding language models'(lack of) theory of mind: A plug-and-play multi-character belief tracker. *arXiv preprint arXiv:2306.00924*.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2023. Clever hans or neural theory of mind? stress testing social reasoning in large language models. *arXiv preprint arXiv:2305.14763*.

Haojun Shi, Suyu Ye, Xinyu Fang, Chuanyang Jin, Leyla Isik, Yen-Ling Kuo, and Tianmin Shu. 2024. Muma-tom: Multi-modal multi-agent theory of mind. *arXiv preprint arXiv:2408.12574*.

Tianmin Shu, Abhishek Bhandwaldar, Chuang Gan, Kevin Smith, Shari Liu, Dan Gutfreund, Elizabeth Spelke, Joshua Tenenbaum, and Tomer Ullman. 2021. Agent: A benchmark for core psychological reasoning. In *International conference on machine learning*, pages 9614–9625. PMLR.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

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

Tomer Ullman, Chris Baker, Owen Macindoe, Owain Evans, Noah Goodman, and Joshua Tenenbaum. 2009. Help or hinder: Bayesian models of social goal inference. *Advances in neural information processing systems*, 22.

Eitan Wagner, Nitay Alon, Joseph M Barnby, and Omri Abend. 2024. Mind your theory: Theory of mind goes deeper than reasoning. *arXiv preprint arXiv:2412.13631*.

Yanming Wan, Jiayuan Mao, and Josh Tenenbaum. 2022. Handmethat: Human-robot communication in physical and social environments. *Advances in Neural Information Processing Systems*, 35:12014–12026.

Qiaosi Wang, Koustuv Saha, Eric Gregori, David Joyner, and Ashok Goel. 2021. Towards mutual theory of mind in human-ai interaction: How language reflects what students perceive about a virtual teaching assistant. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–14.

Ruocheng Wang, Eric Zelikman, Gabriel Poesia, Yewen Pu, Nick Haber, and Noah D Goodman. 2023. Hypothesis search: Inductive reasoning with language models. *arXiv preprint arXiv:2309.05660*.

Alex Wilf, Sihyun Shawn Lee, Paul Pu Liang, and Louis-Philippe Morency. 2023. Think twice: Perspective-taking improves large language models' theory-of-mind capabilities. *arXiv preprint arXiv:2311.10227*.

Heinz Wimmer and Josef Perner. 1983. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128.

Lance Ying, Kunal Jha, Shivam Aarya, Joshua B Tenenbaum, Antonio Torralba, and Tianmin Shu. 2024. GOMA: Proactive embodied cooperative communication via goal-oriented mental alignment. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.

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*, 33:19238–19250.

Tan Zhi-Xuan, Lance Ying, Vikash Mansinghka, and Joshua B Tenenbaum. 2024. Pragmatic instruction following and goal assistance via cooperative language-guided inverse planning. *arXiv preprint arXiv:2402.17930*.

A More Results

A.1 Qualitative Results

Among general methods, AutoToM achieves state-of-the-art results across all benchmarks. We provide two qualitative examples to illustrate the effect of variable adjustment (example 1) and timestep adjustment (example 2). These examples also demonstrate the interpretability of *AutoToM*, as the constructed model offers us insights into how the method is modeling the agent behavior for the inference.

Example 1: BigToM (Backward Belief Inference)

> **Story:** Kavya is a florist in a vibrant Indian market. Kavya wants to create a beautiful bouquet of fresh roses for a customer's anniversary celebration. Kavya sees a batch of roses in her shop that appear to be fresh and vibrant. Unbeknownst to her a mischievous monkey sneaks into the shop and nibbles on the rose petals leaving them damaged and unsuitable for the bouquet. Kavya starts arranging the bouquet using the roses she initially saw.
>
> **Question: Does Kavya believe the roses are fresh and perfect for the bouquet or damaged by the monkey?**
> (a) Kavya believes the roses are fresh and perfect for the bouquet. **(Correct Answer)**
> (b) Kavya believes the roses are damaged by the monkey.

Variables in the Initial Model Proposal: State, Observation, Belief

Since the scenario involves only one timestep, a single model suffices. In the initial model, the state of the world indicates that the flowers are damaged after the monkey nibbles on them. However, it remains unclear whether Kavya observes the true condition of the flowers. The model lacks crucial information about Kavya's actions, which are observable and influenced by her beliefs about the flowers' state. These actions can help infer her true belief. Initially, the probability that Kavya believes the flowers are fresh is moderate, $P(\text{Kavya believes the roses are fresh and perfect for the bouquet}|X^1) = 0.50$. Without variable adjustment, the model cannot answer the question.

Variables in the Adjusted Model: State, Observation, Belief, Action, Goal

For the initial model, the reward is $R(M, q) = -H(P(q|X^{t_s:t})) = -0.693$ and the model cost is $C(M) = \alpha|M| = 0.04$, resulting in a utility $U(M, q) = -0.733$, which does not exceed the utility threshold $U_{\text{min}} = -0.693$. To address the insufficiency of the initial model's utility relative to our termination threshold, we propose an enhanced model incorporating state,

observation, belief, action, and goal. In this revised model, Kavya's actions—specifically arranging the bouquet using the roses—align with her goal of creating a beautiful bouquet. These observations allow us to infer with high probability that Kavya believes the roses are fresh and suitable for the bouquet, increasing the belief probability to $P(\text{Kavya believes the roses are fresh and perfect for the bouquet}|X^1) = 0.97$. With this revised model, the reward is $R(M, q) = -H(P(q|X^{t_s:t})) = -0.135$ and the model cost is $C(M) = \alpha|M| = 0.06$, resulting in a utility $U(M, q) = -0.195$, which exceeds our utility threshold $U_{\text{min}} = -0.693$. Based on the adjusted model, *AutoToM* can confidently determine the correct answer: (a) Kavya believes the roses are fresh and perfect for the bouquet.

Example 2: MMToM-QA (Belief Inference)

Video input:

What's inside the apartment: The apartment consists of a bedroom, kitchen, living room, and bathroom. In the bedroom, there is a coffee table and a desk. The kitchen is equipped with four cabinets, a fridge, a kitchen table, a microwave, and a stove. The 3rd kitchen cabinet from the left houses a water glass and a dish bowl. Inside the fridge, there are two apples, a salmon, a plate, and a dish bowl. The 2nd kitchen cabinet from the left contains a water glass, a chips, a condiment bottle, and a dish bowl. The 1st kitchen cabinet from the left holds a wineglass, a wine, and a condiment bottle. The microwave contains a salmon, and there is a cupcake in the stove. The 4th kitchen cabinet from the left has a plate. The living room features a cabinet, a sofa, a coffee table, and a desk. Inside the cabinet, there are two apples and four books. A plate and a remote control are placed on the coffee table. The bathroom is furnished with a bathroom cabinet, which is currently empty.

Actions taken by Mark: Mark is situated in the bathroom. He proceeds towards the kitchen, making his way to the stove. He opens and then closes the stove. Subsequently, he strides towards the 4th kitchen cabinet, opens it, and then shuts it. He then moves to the 2nd kitchen cabinet, opens and closes it, before doing the same with the 3rd kitchen cabinet. Finally, he heads towards the 1st kitchen cabinet, opens and closes it, and is about to open the microwave.

Question: If Mark has been trying to get a salmon, which one of the following statements is more likely to be true?
(a) Mark thinks that the salmon is not inside the microwave.
(b) Mark thinks that the salmon is inside the microwave. **(Correct Answer)**

In this problem, we first fuse the information from text and video following Jin et al. (2024). The fused information is structured into 23 timesteps, each corresponding to an action of Mark at the time. We then propose the initial model: State, Observation, Belief, Action, Goal.

Without timestep adjustment. Bayesian inference must be performed sequentially from the first timestep, even though most actions do not contribute to answering the final question. The model will compute across all timesteps, while the most informative action is actually the last one: if Mark wants to get a salmon but does not believe there is one inside the microwave, he will not open it.

With timestep adjustment. We begin inference from the last timestep, where the action likelihood $P(a|b, g)$ is low when $b =$ *Mark thinks that the salmon is not inside the microwave*, and high when $b =$ *Mark thinks that the salmon is inside the microwave*. After performing inference at the last timestep, the belief probabilities corresponding to the choices are 0.998 and 0.002. The reward is given by $R(M, q) = -H(P(q|X^{t_s:t})) = -0.014$, while the model cost is $C(M) = \alpha|M| = 0.06$. This results in a utility of $U(M, q) = -0.074$, which exceeds the threshold $U_{\text{min}} = -0.693$, allowing our model to determine the final answer without considering earlier timesteps.

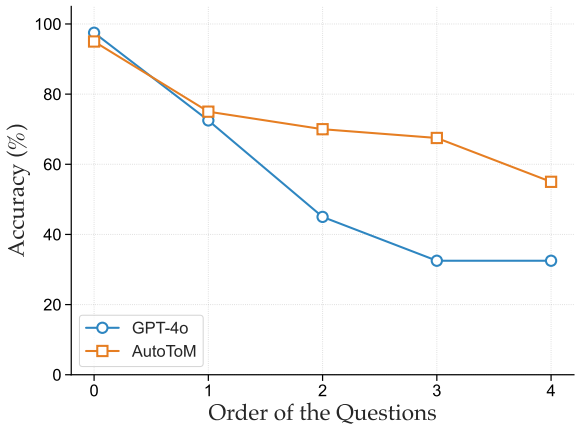

Figure 6: Comparison of accuracy between *AutoToM* and GPT-4o on the HiToM dataset across different reasoning orders. Order 0 refers to questions about an object's actual location; order 1 questions are about an agent's belief about an object's location; order 2 involves questions about an agent's belief regarding another agent's belief, and so forth.

A.2 Results for Higher Order Inference

Higher-order Theory of Mind (ToM) involves recursive reasoning about others' mental states across multiple levels. The Hi-ToM benchmark (He et al., 2023) includes questions ranging from Order 0, which involves no agents and asks about the actual location of objects, up to Order 4, which requires recursive reasoning among four agents. Figure 6 compares the performance of GPT-4o and *AutoToM* across these different question orders. While GPT-4o experiences a significant decline in accuracy as the ToM order increases, *AutoToM* maintains a smaller performance drop and achieves substantially higher accuracy on higher-order questions. This demonstrates that our model-based approach is more robust and scalable, effectively handling complex scenarios involving multiple agents and various levels of recursive reasoning.

A.3 Full Results of the Ablation Study

Table 2 shows the performance of ablated methods compared to the full *AutoToM* method on all benchmarks.

In Table 3 and 4, we compare the ablated methods and the full model on the averaged number of tokens per question (in thousands) and the averaged number of API calls at inference per question.

A.4 Per-type Accuracy on All Benchmarks

In Tables 5 - 9, we present the results of *AutoToM* and baselines on each question type of all benchmarks. Here we compare general methods that can be applied to all benchmarks.

B *AutoToM* Implementation Details

B.1 Variable Adjustments

Table 10 summarizes possible variable adjustments at each timestep.

B.2 Automated Bayesian Inverse Planning

Hypothesis Sampling. At each timestep, hypotheses for the latent variables are generated using a Large Language Model (LLM) as the backend, guided by the observed variables. Specifically, when the state is not explicitly provided, the LLM acts as a world model, tracking state changes in the story based on the previous state and current actions. For an agent's observation, the LLM is prompted to adopt the perspective of a character, simulating what that character might see, know, or hear in the given environment (e.g., inside a closed room). If no new observation is available at a specific timestep, we neither generate new observations nor update the belief. Additionally, the LLM proposes plausible hypotheses for the agent's belief and goal based on the available information.

Hypothesis reduction. We examine all local conditional probabilities involving a single uncertain variable with multiple hypotheses and eliminate those hypotheses that result in significantly low likelihood values. For example, in $P(o^t \mid s^t)$, where s^t represents a determined state, any observation hypothesis that yields a low likelihood for this term is discarded. This approach reduces the computational cost of estimating $P(b^t \mid o^t, b^{t-1})$. Similarly, the same principle is applied to $P(a^t \mid b^t, g^t)$ and $P(u^t \mid b^t, g^t)$, where unlikely belief hypotheses are removed to further reduce computational complexity.

B.3 Automated Model Discovery

When exploring different models during the model discovery, *AutoToM* can reuse the hypothesis proposals of variables and local conditionals from previously computed models to avoid repeated computation.

We configure the hyperparameters in Automated Model Discovery as follows: $\alpha = 0.02$, $U_{\text{min}} = -0.693$.

Method	ToMi	BigToM	MMToM-QA	MuMA-ToM	Hi-ToM	All
w/o hypo. reduction	87.60	86.17	75.83	81.67	69.50	80.15
w/ POMDP	76.00	86.50	79.83	50.78	67.00	72.02
w/o variable adj.	85.80	78.25	76.17	77.89	66.50	76.92
w/ last timestep	68.40	77.83	74.33	78.33	44.50	68.68
w/ all timesteps	86.00	79.09	76.50	79.33	69.00	77.98
AutoToM	88.30	86.92	75.50	81.44	72.50	80.93

Table 2: Results of ablated methods compared to the full *AutoToM* method.

Method	ToMi	BigToM	MMToM-QA	MuMA-ToM	Hi-ToM	All
w/o hypo. reduction	15.8	6.8	19.2	24.4	20.4	17.3
w/ POMDP	14.9	5.5	15.6	20.0	18.8	15.0
w/o variable adj.	8.5	6.1	16.4	14.0	10.0	11.0
w/ last timestep	7.8	6.1	6.4	11.6	4.0	7.2
w/ all timesteps	14.2	7.7	57.2	16.4	12.4	21.6
AutoToM	9.8	6.5	14.4	13.6	12.0	11.3

Table 3: Comparison of ablated models and the full model on the averaged number of tokens per question (in thousands). Lower is better.

Method	ToMi	BigToM	MMToM-QA	MuMA-ToM	Hi-ToM	All
w/o hypo. reduction	38.91	13.99	45.97	70.73	72.58	48.44
w/ POMDP	36.25	8.32	41.18	42.10	51.73	35.92
w/o variable adj.	22.91	12.99	35.46	35.76	29.81	27.39
w/ last timestep	21.60	12.76	12.75	28.39	9.39	16.98
w/ all timesteps	39.83	15.95	116.81	43.25	36.27	50.42
AutoToM	32.23	13.81	31.36	35.08	36.45	29.79

Table 4: Comparison of ablated models and the full model on the averaged number of API calls at inference per question. Lower is better.

Question Type	First order	Second order	Reality	Memory	All
Llama 3.1 70B	73.75	56.25	100.00	100.00	72.00
Gemini 2.0 Flash	58.50	58.25	100.00	100.00	66.70
Gemini 2.0 Pro	75.00	54.75	100.00	100.00	71.90
GPT-4o	80.25	62.25	100.00	100.00	77.00
SimToM	84.75	65.00	100.00	100.00	79.90
AutoToM	95.00	77.50	93.00	100.00	88.30

Table 5: Detailed accuracy for ToMi.

Question Type	Forward TB	Forward FB	Backward TB	Backward FB	All
Llama 3.1 70B	93.75	81.00	57.00	60.50	77.83
Gemini 2.0 Flash	94.25	87.50	77.50	51.00	82.00
Gemini 2.0 Pro	96.00	93.75	70.00	68.50	86.33
GPT-4o	96.00	88.50	63.50	62.00	82.42
SimToM	92.50	90.00	25.00	75.00	77.50
AutoToM	91.25	93.75	73.00	78.50	86.92

Table 6: Detailed accuracy for BigToM.

B.4 Recursive Reasoning

Interactive Partially Observable Markov Decision Process (I-POMDP) extends POMDP to multi-agent settings by introducing the concept of interactive states, which include agent models into the state space to capture the recursive reasoning process (Gmytrasiewicz and Doshi, 2005). We denote $is_{i,l}$ as the interactive state of agent i at level l. For two agents i and j, where agent i is interacting with agent j, the interactive states at each level are

Question Type	Belief	Goal	All
Llama 3.1 70B	51.33	36.33	43.83
Gemini 2.0 Flash	62.67	33.33	48.00
Gemini 2.0 Pro	57.00	44.67	50.84
GPT-4o	55.67	32.33	44.00
SimToM	75.67	26.33	51.00
AutoToM	88.67	62.33	75.50

Table 7: Detailed accuracy for MMToM-QA.

Question Type	Belief	Goal	Belief of Goal	All
Llama 3.1 70B	68.67	51.33	47.33	55.78
Gemini 2.0 Flash	68.33	50.67	47.00	55.33
Gemini 2.0 Pro	63.00	66.67	57.00	62.22
GPT-4o	85.33	57.00	48.33	63.55
SimToM	54.60	43.50	44.80	47.63
AutoToM	88.33	77.00	79.00	81.44

Table 8: Detailed accuracy for MuMA-ToM.

Question Type	Order 0	Order 1	Order 2	Order 3	Order 4	All
Llama 3.1 70B	65.00	47.50	22.50	20.00	20.00	35.00
Gemini 2.0 Flash	95.00	70.00	50.00	27.50	20.00	52.50
Gemini 2.0 Pro	100.00	62.50	50.00	37.50	37.50	57.50
GPT-4o	92.50	65.00	40.00	27.50	25.00	50.00
SimToM	100	77.50	60.00	60.00	57.50	71.00
AutoToM	95.00	75.00	70.00	67.50	55.00	72.50

Table 9: Detailed accuracy for HiToM.

New Var.	Before	After
Goal	$P(a^t \mid s^t)$	$P(a^t \mid s^t, g)P(g)$
	$P(a^t \mid b^t)$	$P(a^t \mid b^t, g)P(g)$
	$P(a^t)$	$P(a^t \mid s^t, g)P(g)$
	$P(a^t)$	$P(a^t \mid b^t, g)P(g)$
Belief	$P(a^t \mid s^t)$	$P(a^t \mid b^t)P(b^t \mid s^t, b^{t-1})$
	$P(a^t \mid s^t, g)$	$P(a^t \mid b^t, g)P(b^t \mid s^t, b^{t-1})$
Observ.	$P(b^t \mid s^t, b^{t-1})$	$P(b^t \mid o^t, b^{t-1})P(o^t \mid s^t)$
Int. State	$b(s^t)$	$b(is^t)$

Table 10: Potential variable adjustments, including introducing goal, belief, observation, and interactive state (for high-order ToM). We show the corresponding local conditionals before and after introducing the new variables.

defined as:

- **Level 0:** $is_{i,1} = s$
- **Level 1:** $is_{i,1} = (s, b_{j,0}, g_j)$ where $b_{j,0}$ is a distribution over j's interactive state at level 0, $is_{j,0}$
- ...

The framework provides a generative model for agents: given agent i's belief of interactive state $b(is_{i,l})$, its action policy will be $\pi(a_i|is_{i,l}, g_i)$, and its utterance policy will be $\pi(u_i|is_{i,l}, g_i)$.

In our implementation, we sample one possible state based on $b(s)$ at level l to approximate the state at level $l-1$ as imagined by the agent at level l. We can recursively apply this process until reaching level 0. Based on the state sampled for level 0, we can then conduct the typical automated BIP based on the model structure at that level. This approach can be conveniently applied to arbitrary levels of recursive reasoning, allowing us to answer higher-order Theory of Mind questions using the same method.

B.5 BToM Model Space

To apply Bayesian Inverse Planning (BIP) across various scenarios, we define the mental variables and their causal relationships with agent behavior using a family of Bayesian Theory of Mind (BToM) models. These models accommodate different levels of complexity in how agents behave and reason about their environment.

At each timestep t, the observable variables are represented by:

$$X^t = \{x_i^t\}_{i \in N_X}, \text{ where } N_X = \{s^t, a^t, u^t\}$$

Here, the state s^t always appear in X^t, while either a^t (action) or u^t (utterance) is included at timestep t, depending on whether physical motion or verbal communication is presented. In some cases, a^t is only used to update the state and does not affect the inference of beliefs or goals, while in other scenarios it can be crucial for inferring hidden mental states (e.g., an agent's belief or goal).

The latent variables are denoted by

$$V^t = \{v_i^t\}_{i \in N_V}, \text{where } N_V = \{o^t, b^t, g^t\}$$

Here, the observation o^t is only included when the agent's belief b^t is part of the model, as it updates b^t. The goal g^t is included only if it influences action and is relevant to inference. In cases of higher-order recursive reasoning among multiple agents, the belief over the state $b^t(s^t)$ extends to belief over an interactive state $b^t(is^t)$.

Combining these choices at each timestep yields a model space with 30 possible configurations:

- Action/Utterance: which one is included (2 options).
- Belief/Observation: no belief, belief of state, belief of interactive state, belief of state, or belief of interactive state + observation (5 options).
- Action(Utterance)/Goal: no goal (action(utterance) irrelevant), action(utterance) only, or action(utterance) + goal (3 options).

Over a time interval from t_s to t, this scales to 30^{t-t_s+1} possible models.

Examples. In addition to the Markov Decision Process (MDP), Partially Observable Markov Decision Process (POMDP), and Interactive POMDP (I-POMDP) models introduced in Section 3.1, we present additional examples of models from the BToM model space:

- Observation Update Model: Used in the ToMi benchmark (see Figure 2), this model focuses on how observations update beliefs. Actions are present but only serve to update states and are irrelevant to the inference questions. This model is well-suited for passive scenarios where the focus is on understanding how hidden states produce observable evidence and how the agent updates its beliefs about the world.
- POMDP Variant without Goal: A partially observable scenario in which goals are trivial or irrelevant. This variant emphasizes how partial observability affects belief formation and action selection, without explicit goal-driven behavior.

C Baseline Implementation Details

For the baselines, we use `gpt-4o-2024-08-06` for GPT-4o, `meta-llama/Llama-3.1-70B-Instruct` from Hugging Face for Llama 3.1 70B, `gemini-2.0-flash` for Gemini 2.0 Flash, and `gemini-2.0-pro-exp-02-05` for Gemini 2.0 Pro. Among the ToM prompting for LLM benchmarks previously tested on the BigToM dataset, e.g., TimeToM and SimToM, they only tested the subset of the entire dataset with questions for forward action and forward belief and did not test on backward belief questions. With the available SimToM code, we tested it on the full BigToM dataset with GPT-4o, while TimeToM does not have its code available.

SymbolicToM maps out the agents' beliefs throughout stories of different levels of reasoning via symbolic graphs. However, the construction of these graphs is specifically designed for the ToMi dataset, where there are fixed actions and sentence formats in the story. Thus it is difficult to generalize to more open-ended scenarios (e.g., BigToM) or stories with multiple agents acting simultaneously (e.g., Hi-ToM). Therefore, we can only evaluate SymbolicToM on ToMi (tested with GPT-4o on the full dataset), for which it was designed.

TimeToM is not open-source. We rely on its self-reported accuracy on ToMi. However, since it was only evaluated on a subset of BigToM with forward inference questions, its accuracy on the full BigToM benchmark remains unknown. Similarly, PercepToM is not open-source, and we rely on its self-reported accuracy on ToMi.

BIP-ALM and LIMP are both models that combine BIP and LLMs to solve ToM problems. BIP-ALM manually defines symbolic representations of observable and latent variables and assumes POMDP. LIMP is designed to only solve two-level reasoning problems. It uses natural language to represent variables. Both methods assume that the goals are about finding an object and the beliefs are about the locations of that object in a household environment.

Benchmark	Agent number	Tested concepts	Size	Modality	Communication	Generation	Evaluation
ToMi (Le et al., 2019)	Multi agents	First & Second Order belief, Reality, Memory	1000	Text	No	Templates	Multiple choice Q&A
BigToM (Gandhi et al., 2024)	Single agent	Belief, Action	1200	Text	No	Procedural generation	Question answering
MMTOM-QA (Jin et al., 2024)	Single agent	Belief & Goal	600	Text & Video	No	Procedural generation	Multiple choice Q&A
MuMA-ToM (Shi et al., 2024)	Multi agents	Belief, social goal and belief of other's goal	900	Text & Video	Yes	Procedural generation	Multiple choice Q&A
Hi-ToM (He et al., 2023)	Multi agents	High-order beliefs	200	Text	Yes	Procedural Generation	Multiple choice Q&A

Table 11: Summary of the ToM benchmarks used in the experiments.

D Benchmark Details

In our evaluation, we test *AutoToM* on BigToM (Gandhi et al., 2024), MMToM-QA (Jin et al., 2024), MuMA-ToM (Shi et al., 2024), ToMi (Le et al., 2019) and Hi-ToM (He et al., 2023). For ToMi, we use the ToMi dataset that has disambiguated container locations in the story and correctly labeled order of reasoning (Arodi and Cheung, 2021; Sap et al., 2022). For Hi-ToM, we choose the length 1 subset consisting of 200 questions across all orders (0-4) due to the high cost of testing the full dataset.

Table 11 summarizes the benchmarks used to evaluate *AutoToM* against baselines, detailing key features such as test concepts, input modalities, and the number of agents. The results demonstrate that *AutoToM* operates across diverse contexts, infers any mental state, reasons about any number of agents, and supports any level of recursive reasoning.

E Prompts used in *AutoToM*

E.1 Information Extraction

We use the following prompts to extract information for each variable in a given question.

Identifying the main agent

Find the name of the character that we need to infer about in the question and choices. Only output the name. Do not answer the question.

Question: [Question]
Choices: [Choices]
Character name:

Identifying all the agents

Extract the names of all the characters from the story and question. Provide only the names or roles, without any additional information. Do not answer the question. Your response should be a list containing the names, like ["name1", "name2"].

Story: [Story]
Response:

Identifying the mental variable to be inferred

Choose the variable that best summarizes the information about the differences that the choices contain. Only output the variable.

Variables include: [Variables]
Choices: [Choices]
Variable:

Identifying extra information in the question

If there is any assumed information in the question given (a conditional clause starting with specific words like "if" is contained), rewrite it as a declarative sentence. Do not include any questions in the extra informa-

tion. Do not make up details for the information. Use the original wording.
Otherwise, output "NONE".

Question: [Question]
Extra Information:

Extracting actions of the main agent

Extract the actions of [Inferred_agent] in the story verbatim without changing any of the original words, pluralizing the words, adding in [Inferred_agent] or any other name, replacing any of the words, replacing pronouns with names or replacing any names with pronouns. Actions of [Inferred_agent] are defined as events that will change the world state, e.g., [Inferred_agent] moving to a new location is an action but [Inferred_agent] being at a location is not an action. If [Inferred_agent] says something, the whole sentence (with "replied", "said") is seen as an action.
Do not change the names of any of the agents, if there is not a name and only a pronoun then just leave the pronoun. There can be more than one agent or more than just the inferred agent.
If there are multiple actions in a sentence then they should be extracted as one single action, without changing any of the original words, such as pluralizing the words, replacing any of the words, replacing pronouns with names, or replacing any names with pronouns, and do not add any words.
Do not insert actions, pronouns, or other words that are not explicitly stated in the text. Do not separate the objects in the same action.
Do not add any pronouns. Keep the commas, if any.
Only actions that have already occurred at the time can be considered clearly stated. Again, only extract actions performed by [Inferred_agent].
The output format should be: ["aaa.", "bbb.", ...]. Output only this list.

Story: [Story]
Extraction:

Extracting actions

Determine if [Character]'s action(s) is clearly stated in the story.
The action(s) cannot be the character's inner thoughts.
Only actions of [Character] that have already occurred, or are currently taking place can be considered clearly stated.
If it's more like [Character]'s desire or goal, it does not count as an action. [Character]'s utterance is considered as an action (include the verb like "said" or "replied" in the evidence sentence, if any). Do not change any of the original wording.
Answer in the form of a list. The first element of the list contains the option A or B. A means clearly stated, and B means not clearly stated.
If the answer is A, include sentence(s) from the original story that serves as evidence, and place it in the second element of the list, without any kind of formatting. Note that there could be multiple action sentences.
Otherwise, the second element can be an empty string. Do not write anything else.
Example 1: ["A", "evidence sentence."]
Example 2: ["B", ""]

Story: [Story]
Answer:

Extracting beliefs

Determine if the belief of [Character] is clearly stated in the story.
Usually, belief is one's understanding of the state of the world or the state of others. A subjective attitude towards things does not count as belief. An action or utterance of the agent does not count as a belief. Words like "know" or "believe" could be hints for belief.
Answer in the form of a list. The first element of the list contains the option A or B. A means clearly stated, and B means not clearly stated.
If the answer is A, include sentence(s) from the original story that serves as evidence, and place it in the second element of the list, without any kind of formatting.

Otherwise, the second element can be an empty string. Do not write anything else.
Example 1: [“A”, “evidence sentence.”]
Example 2: [“B”, “”]

Story: [Story]
Answer:

Extracting goals

Determine if the goal of [Character] is clearly stated in the story.
Usually, goals refer to a person’s goals or intentions regarding a particular event. Moreover, a sentence that shows a person has been trying to do something, or summarizes their efforts of doing something should always be considered a goal. Helping others to achieve their goals also counts as a person’s goal.
Answer in the form of a list. The first element of the list contains the option A or B. A means clearly stated, and B means not clearly stated.
If the answer is A, include sentence(s) from the original story that serves as evidence, and place it in the second element of the list, without any kind of formatting.
Otherwise, the second element can be an empty string. Do not write anything else.
Example 1: [“A”, “evidence sentence.”]
Example 2: [“B”, “”]

Story: [Story]
Answer:

Extracting observations

Determine if the observation of [Character] is clearly stated in the story.
Observation refers to the main character’s perception of an event; it is only considered clearly stated when the protagonist’s perception is explicitly mentioned, like if they visually see something, visually notice something, or hear something, or any other state that can be perceived by the agent with but not limited to their 5 senses.
A character’s utterance does not mean that their observation is clearly stated, because they might lie.
Answer in the form of a list. The first element of the list contains the option A or B. A means clearly stated, and B means not clearly stated.
If the answer is A, include sentence(s) from the original story that serves as evidence, and place it in the second element of the list, without any kind of formatting.
Otherwise, the second element can be an empty string. Do not write anything else.
Example 1: [“A”, “evidence sentence.”]
Example 2: [“B”, “”]

Story: [Story]
Answer:

Extracting states

Determine if the story contains the objective state(s) of an object or an event.
State refers to the physical condition of something or the state of the world.
No actions of agents should be involved in the state but it can be the result of an action of an agent. For example, “A entered B” is not a state, while “A is in B” is a state.
An objective state statement should not include personal perspectives but should be objective. If a person’s perception is involved, it is no longer considered an objective state.
Answer in the form of a list. The first element of the list contains the option A or B. A means clearly stated, and B means not clearly stated.
If the answer is A, include sentence(s) from the original story that serves as evidence, and place it in the second element of the list, without any kind of formatting.
If there are multiple sentences, include them all in the second element of the list.
Otherwise, the second element can be an empty string. Do not write anything else.
Example 1: [“A”, “evidence sentence(s).”]
Example 2: [“B”, “”]

Story: [Story]
Answer:

E.2 Hypothesis Sampling

We use the following prompts to sample hypotheses for the latent variables in the BToM models.

Sampling beliefs

Propose [num] hypotheses for the belief of [Character] in the story aligned with the context of: [Context]. Make sure that it is not any of the hypotheses in [Wrong Hypotheses], if it is then propose new hypotheses that are very different.
It should be related to [Information] and the context described above.
The hypotheses do not require reasoning or consideration of whether they are likely to occur. The only limitation is that they must be relevant to the information already provided. You cannot return nothing. Usually, belief is one's view or perspective on a matter, and it represents an understanding of the state of the world or the state of others. The emotional attitudes toward a specific thing do not count as belief. Do not state any reason for the hypotheses. Do not contain any form of explanation in the hypotheses. Output a list of hypotheses of length [num] in the following form: ["aaa.", "bbb.", ...]

Context: [Context]
Belief Hypotheses:

Sampling goals

Propose [num] hypotheses for the goal of [Character].
The goal refers to [Character]'s intentions.
Do not provide any explanation for the hypotheses. Do not propose any sentence that's not depicting the goal, like the action or belief of [Character].
The wording for hypotheses cannot be speculative.
The proposed goal does not have to be too specific, e.g., Andy wants to help others; Andy wants to hinder others; Andy is indifferent towards other's goals, etc.
Given information: [Information]
Ensure that the hypotheses align with the given information perfectly. It means that the proposed [Character]'s goal matches what's contained in the information.
Output the hypotheses in the following form: ["aaa."]

Goal Hypotheses: []

Sampling observations

Propose [num] hypotheses for [Character]'s observation of the world.
The observation refers to [Character]'s current perception of events or the world state. It is only considered clearly stated when [Character]'s perception is explicitly mentioned, like if [Character] sees something or perceives something through other senses. Do not be speculative.
Do not provide any explanation for the hypotheses. Do not propose any sentence that's not depicting the observation, like the action or belief of [Character].
The wording for hypotheses cannot be speculative.
If the information contains "not", make sure the verb for perception (e.g., "see", 'perceives') goes before "not" in the hypotheses. e.g., use 'sees that A is not in B' instead of 'does not see that A is in B' Otherwise, do not include "not" in your hypotheses, and make sure the verb for perception goes first, e.g., 'sees that A is in B'.
Given information: [Information]
Ensure that the hypotheses align with the given information perfectly. It means that when the person has the observation the person will act according to the given information.
First, list all entities in the given information. Then, formulate hypotheses using all entities. Make sure the hypothesis starts with [Character].
Output the hypotheses in the following form: ["aaa."]

Observation Hypotheses: []

E.3 Likelihood Estimation

We use the following prompts to estimate the likelihood of different variables.

Estimating the likelihood of the observation given the state

Determine if the statement is likely, and respond with only either A or B.
State: {state}
Here is a statement of {inf_agent}'s current observation. Only evaluate current observation of {inf_agent} based on the state. Do not imagine anything else. Think about {inf_agent}'s location. {inf_agent} is quite likely to observe all objects and events in {inf_agent}'s location, and is unlikely to observe states in another location. If {inf_agent} does not appear in the state, {inf_agent} can't observe anything. Note that the statement has to be precise in wording to be likely. For example, the treasure chest and container are different in wording and they're different objects.

Determine if the following statement is likely: {statement}
A) Likely.
B) Unlikely.

Estimating the likelihood of the action given the goal and belief and belief of goal

Determine if the statement is likely, and respond with only either A or B.
{inf_agent}'s goal: {goal}
{inf_agent}'s belief: {belief}
{inf_agent}'s belief of other's goal: {belief of goal}
{inf_agent}'s action: {action}
When {inf_agent} wants to help, {inf_agent} is likely to bring an object to other's desired location, and unlikely to grab an object away from other's desired location.
When {inf_agent} wants to hinder, {inf_agent} is likely to grab an object away from other's desired location, and unlikely to bring an object to other's desired location.
When {inf_agent} doesn't know other's goal, {inf_agent} is likely to act according to {inf_agent}'s belief.
If {inf_agent} wants to help and {inf_agent} believes the object is placed at other's desired location, it's unlikely {inf_agent} will move the object.
If {inf_agent}'s goal, {inf_agent}'s belief of goal, and {inf_agent}'s action do not align in any way, the action is unlikely.

Determine if {inf_agent}'s action is likely.
A) Likely.
B) Unlikely.

Estimating the likelihood of the action given the goal and belief

Determine if the statement is likely, and respond with only either A or B. If it's not certain but it's possible, it's likely.
{inf_agent}'s goal: {goal}
{inf_agent}'s belief: {belief}
Here is a statement of {inf_agent}'s action. Think about {inf_agent}'s goal.
{inf_agent} will perform actions according to {inf_agent}'s belief, and any action that does not align with the belief is very unlikely, except when {inf_agent}'s goal is to hinder or to prevent others. In this case (goal is hindering others) {inf_agent}'s action is only likely when it's different from {inf_agent}'s belief. If {inf_agent}'s mental states contain conditions like "When giving information" and the action is not giving information, it's unlikely.

Determine if the following statement is likely: {statement}
A) Likely.
B) Unlikely.

Estimating the likelihood of the best action among choices given the goal and belief

Determine if the statement is likely, and respond with only either A or B. If it's not certain but it's possible, it's likely.
{inf_agent}'s belief: {belief}
{inf_agent}'s goal: {goal}
If the next immediate actions possible are: {actions}
Determine which immediate action is

most possible given the information about {inf_agent}'s goal and belief.

Determine if the following statement is likely: {action_a} is a better immediate action than {action_b}.
A) Likely.
B) Unlikely.

Estimating the likelihood of the initial belief

Determine if the statement is likely, and respond with only either A or B. If it's not certain but it's possible, it's considered likely.
Here is a statement of the story and {inf_agent}' initial belief.
There is an action that causes the state of the main object to change. Based on {inf_agent}'s observations determine if {inf_agent} perceives the state of the object change.
If it is not clearly stated that {inf_agent} perceives it then we do not assume that {inf_agent} perceived the change of state.
If {inf_agent} perceives this change then it is highly likely that {inf_agent}'s belief aligns with the change of state of the object.
If {inf_agent} does not perceive this change or if it is unknown if {inf_agent} perceives this change then it is highly likely that {inf_agent}'s belief does not align with the change of state of the object.
Story: {story}
Think about the state of the world and others actions. {inf_agent}' belief can change throughout time through other's actions and what {inf_agent} can observe. It is also important to think about if {inf_agent} can observe other's actions. If {inf_agent} can observe the same then their belief will change and if not then their belief will remain constant. Use this to determine {inf_agent}'s beliefs.

Determine if the following statement is likely: {statement}
A) Likely.
B) Unlikely.

Estimating the likelihood of the belief given the observation and previous belief

Determine if the statement is likely, respond with only either A or B.
{inf_agent}'s previous belief: {previous_belief}
{inf_agent}'s observation: {observation}
Here is a statement of {inf_agent}'s current belief. If {inf_agent}'s current belief is not aligned with {inf_agent}'s observation, it is very unlikely.

Determine if the following statement is likely: {statement}
A) Likely.
B) Unlikely.

Estimating the likelihood of the belief given the state and previous belief

Determine if the statement is likely, respond with only either A or B.
{inf_agent}'s previous belief: {belief}
State: {state}
Here is a statement of {inf_agent}'s current belief. If {inf_agent}'s current belief is not aligned with the state, it is very unlikely.

Determine if the following statement is likely: {statement}
A) Likely.
B) Unlikely.

Estimating the likelihood of the utterance

Determine if {inf_agent}'s utterance is likely, and respond with only either A or B.
{inf_agent}'s belief: {belief}
{inf_agent}'s goal: {goal}
{inf_agent}'s utterance: {utterance}
When {inf_agent}'s goal is to help others, {inf_agent}'s utterance is likely when it strictly reflects {inf_agent}'s belief, and unlikely if it does not reflect {inf_agent}'s belief.
When {inf_agent}'s goal is to hinder or to prevent others from achieving their goals, {inf_agent}'s utterance is likely when it's different from {inf_agent}'s belief, and unlikely if it reflects {inf_agent}'s belief.

Determine if {inf_agent}'s utterance is likely.
A) Likely.
B) Unlikely.

E.4 Initial Model Proposal

We use the following prompts to propose an initial model for a question and determine if the question has higher-order beliefs.

Proposing the initial model

What variables are necessary to solve this question? Please provide the answer without an explanation.
Please select from the following: ["State", "Observation", "Belief", "Action", "Goal"]
State: The true condition of the environment. This should always be included.
Observation: The observed information about the state. Include this when the agent has partial observations of the state.
Belief: The agent's current estimation of the true state is based on the state or its observation.
Action: A move made by the agent, informed by the state or belief. Include this only when it is directly relevant to answering the question.
Goal: The objective the agent is trying to achieve. Include this only if "Action" is included.

Question:{example_question}
Variables: {example_answer}
Question: {question}
Variables:

Determining if the question contains a higher-order belief

Determine whether the question is about a higher-order belief.
A higher-order belief refers to a belief about another person's belief, goal, or action.
It is not a high-order belief if it only asks about one agent's belief.
Please respond with "Yes" or "No".
If the answer is "Yes", the question often ends with "Where does A think that B ...?"
Otherwise, respond "No".

Question: [A story involving several people.] Where will Jack look for the celery?
Higher-order belief: No
Question: [A story involving several people.] Where does Jack think that Chloe searches for the hat?
Higher-order belief: Yes
Question: {question}
Higher-order belief: