Large Language Models Show No Coherence Across Different Theory of Mind Tasks: Evidence From GPT-4o

Anonymous submission

Abstract

Large Language Models (LLMs) have recently shown success across a range of social tasks, raising the question of whether they have a Theory of Mind (ToM). Research into this question has focused on evaluating LLMs against benchmarks, rather than testing for the representations posited by ToM. Using a cognitively-grounded definition of ToM, we develop a new evaluation framework that allows us to test whether LLMs have a mental causal model of other minds (ToM), human-like or not. We find that LLM social reasoning lacks key signatures expected from a causal model of other minds. These findings suggest that the social proficiency observed in LLMs is not the result of a ToM.

Website — Anonymized

Introduction

Large Language Models (LLMs) are not only proficient language users, but also social reasoners. They can infer indirect meanings in language (Hu et al. 2022), make simple moral judgments (Almeida et al. 2024), and plan cooperative behavior (Guo et al. 2024; Shen et al. 2024). In humans, these capacities rely on Theory of Mind (ToM) (Rubio-Fernandez, Berke, and Jara-Ettinger. in press; Young et al. 2007; Ullman et al. 2009), raising the question of whether this capacity has spontaneously emerged in LLMs.

Research into LLM ToM shows conflicting results, with some work showing remarkable successes (Kosinski 2023), and other revealing striking brittleness (Ullman 2023). Here we offer a new proposal for testing LLM ToM that moves away from traditional benchmarking approaches, focusing instead on the defining internal representations that constitute ToM.

In cognitive science, ToM is defined as a causal model of how mental states produce behavior, which we can use to predict action given mental states and invert to infer mental states from action (Gopnik and Meltzoff 1997). In humans, the forward model (mental states to actions) is structured around a principle of rational planning (Gergely and Csibra 2003; Jara-Ettinger et al. 2016), and the inferences (actions to mental states) invert the forward model via Bayesian inference (Baker et al. 2017).

The cognitive definition of ToM reveals two critical considerations. First, there is not one but many ToMs. The causal model used to explain behavior is different in children and adults (Onishi and Baillargeon 2005; Wellman and Liu 2004), it is different between human and non-human primates (Martin and Santos 2016; Rosati, Santos, and Hare 2010), and shows some variability across cultures (Yu and Wellman 2024; Liu et al. 2008). In the same way, LLMs might have their own emergent ToM – one that differs from human ToM and therefore might be missed by benchmarking tests. Second, because action predictions and mental-state inferences result from forwards and backwards outputs of the same causal model, they are fundamentally linked and should be coherent given corresponding inputs. As such, we propose to test for LLM ToM using parametrically varying scenarios to examine the coherence between its action predictions and mental-state inferences.

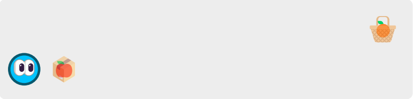

Figure 1: An instance of *ContainerWorld* with an apple in the box and an orange in the basket.

Evaluation Method

Fig. 2 shows our approach. We construct a simple paradigm that allows us to enumerate all the possible beliefs, desires, and world states of an event, and query the LLM for an action prediction – i.e., mapping its forward model (Fig. 2A). We then use the forward model as a likelihood function to infer mental states from action, and compare these expected inferences from ToM to the ones directly produced by the LLM.

Our paradigm, *ContainerWorld*, is shown in Fig. 1. A character always begins next to a closed box, with a covered basket fifty steps away. Each container will have either apples, oranges, or both (apples and oranges) $\mathcal{S} \in \{\text{apples}, \text{oranges}, \text{apples and oranges}\}$. The agent has desires $\mathcal{D} \in \{\text{likes}, \text{dislikes}\}$ towards apples and oranges (excluding the configuration where the agent dislikes both fruits), and beliefs about the contents of each container, $\mathcal{B} \in \{\text{apples}, \text{oranges}, \text{apples and oranges}\}$.

We transcribe *ContainerWorld* into prompts, and query an LLM to predict which container the agent will move to,

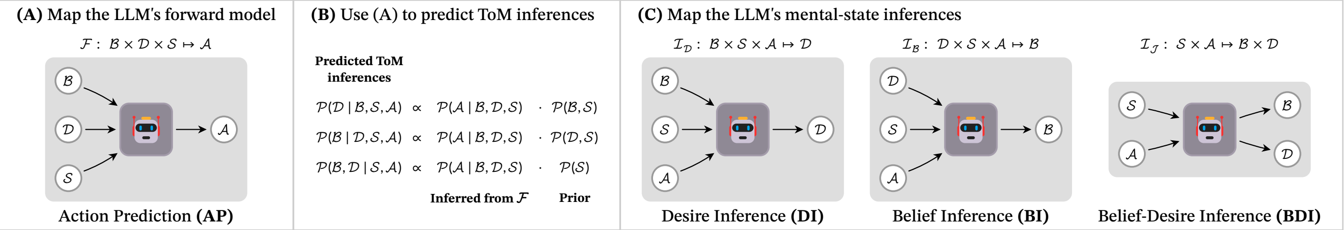
(A) Map the LLM's forward model
$\mathcal{F}: \mathcal{B} \times \mathcal{D} \times \mathcal{S} \mapsto \mathcal{A}$
$\mathcal{B}$
$\mathcal{D}$
$\mathcal{S}$
$\mathcal{A}$
Action Prediction (AP)
(B) Use (A) to predict ToM inferences
Predicted ToM inferences
$\mathcal{P}(\mathcal{D} \mid \mathcal{B},\mathcal{S},\mathcal{A}) \propto \mathcal{P}(\mathcal{A} \mid \mathcal{B},\mathcal{D},\mathcal{S}) \cdot \mathcal{P}(\mathcal{B},\mathcal{S})$
$\mathcal{P}(\mathcal{B} \mid \mathcal{D},\mathcal{S},\mathcal{A}) \propto \mathcal{P}(\mathcal{A} \mid \mathcal{B},\mathcal{D},\mathcal{S}) \cdot \mathcal{P}(\mathcal{D},\mathcal{S})$
$\mathcal{P}(\mathcal{B},\mathcal{D} \mid \mathcal{S},\mathcal{A}) \propto \mathcal{P}(\mathcal{A} \mid \mathcal{B},\mathcal{D},\mathcal{S}) \cdot \mathcal{P}(\mathcal{S})$
Inferred from $\mathcal{F}$
Prior
(C) Map the LLM's mental-state inferences
$\mathcal{I}_{\mathcal{D}}: \mathcal{B} \times \mathcal{S} \times \mathcal{A} \mapsto \mathcal{D}$
$\mathcal{I}_{\mathcal{B}}: \mathcal{D} \times \mathcal{S} \times \mathcal{A} \mapsto \mathcal{B}$
$\mathcal{I}_{\mathcal{J}}: \mathcal{S} \times \mathcal{A} \mapsto \mathcal{B} \times \mathcal{D}$
Desire Inference (DI)
Belief Inference (BI)
Belief-Desire Inference (BDI)

Figure 2: For each pairing of relevant $\mathcal{B}, \mathcal{D}, \mathcal{S}, \mathcal{A}$, we query the LLM for a distribution over predictions (2A) and mental states (2C). Additionally, we compute the predicted inference under a Bayesian inversion, using the distributions provided in (2B).

$\mathcal{A} \in \{\text{box}, \text{basket}\}$, for the full configuration of states, beliefs, and desires ($9 \times 9 \times 3$). We use the distribution over next tokens as the LLM's likelihood of the action.

We then test if an LLM's forward model predicts its mental-state inferences from action (regardless of its agreement with human intuitions). Specifically, we test for prediction-inference agreement across three mental-state inference tasks: desire inference, belief inference, and joint belief-desire inference (Fig. 2C). In each case, we compute the predicted inference under a Bayesian inversion of the forward model (Fig. 2B), take the expected posterior (as humans do; Baker et al. 2017), and compare it to the token likelihood extracted directly from the LLM (Fig. 2C) – the "Bayesian" evaluation. It is also possible that an LLM is relying on forward model expectations to produce inferences, but not in a Bayesian way. We therefore also consider a more generous evaluation metric: a mental-state inference is consistent if, when used as input to the forward model $\mathcal{F}$, it produces the target action to be explained – the "validity" evaluation. This is a generous metric because, for any action, there is a large space of possible inputs that can generate it.

Results

We evaluate our approach using `gpt-4o-2024-05-13` (GPT-4o). In our "Bayesian" evaluation, we expect that a GPT-4o's direct estimates will highly, positively, correlate with its Bayesian inversion – instead, we find that its mental-state inference estimates do not positively correlate with its Bayesian inversion (Fig. 3A). In our "validity" evaluation, we would expect that the forward model $\mathcal{F}$ and each inference model $\mathcal{I}$ would fully agree – instead, we find that GPT-4o's prediction and inference models agree more often than not (Fig. 3B).

To ensure that these results are not because *ContainerWorld* is unusually a challenging domain for ToM in GPT-4o, we constructed a logically equivalent paradigm *MovieWorld* and repeated our evaluation scheme. We find similarly low correlations in our "Bayesian" evaluation (DI: $r = .57$, 95% CI $[.49, .63]$; BI: $r = .03$, 95% CI $[-.05, .12]$; BDI: $r = .13$, 95% CI $[.03, .12]$). In our "validity" evaluation, we find striking agreement in *MovieWorld* (DI: 83.5%; BI: 81.1%; BDI: 88.9%).

The overall low correlations in our "Bayesian" evaluation and high agreements in our "validity" evaluation illustrate that GPT-4o's action predictions (from mental states) are unrelated to its mental-state inferences (from actions). It is possible, however, that GPT-4o does not re-use the forward model for inference, but still learn a global forward and inference model that is context independent. To test this, we evaluated whether GPT-4o produced consistent behavior across the two logically equivalent paradigms, comparing the forward models in *ContainerWorld* to *MovieWorld*, and the inferences in all three tasks. Fig. 3C shows that, despite their equivalence, GPT-4o's behavior shows no consistency across tasks.

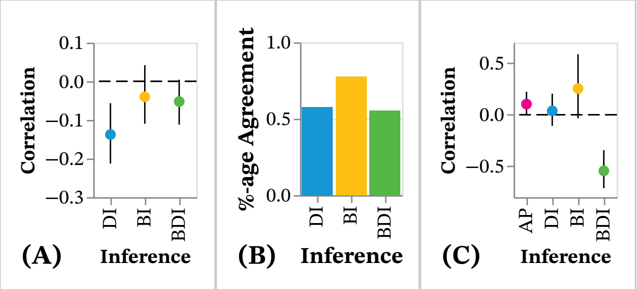

Figure 3: GPT-4o's coherence under the Bayesian approach (3A), percentage of actions produced by the mental-state inference (3B), and coherence in GPT-4o's predictions across all tasks in two logically-equivalent domains (3C).

Discussion

This work makes three contributions. First, we propose a new way to test for LLM ToM that moves away from benchmarking metrics, to testing for the representational signatures of ToM. This approach can differentiate social mimicry (high benchmark performance with no ToM representations) from non-human forms of ToM (low benchmark performance, but internal coherence pointing to ToM representations). Second, we show that GPT-4o lacks coherence between forward and inverse mappings. This contrasts with the representations posited in ToM, which involve a causal model that is used to both predict and interpret others' behavior. Third, we show that GPT-4o's action predictions and mental-state inferences were not consistent across two logically-equivalent tasks. This suggests that GPT-4o lacks a coherent set of agent expectations that transfers across domains. In future work we plan to evaluate other LLMs.

References

Almeida, G. F. C. F.; Nunes, J. L.; Engelmann, N.; Wiegmann, A.; and Araújo, M. d. 2024. Exploring the psychology of LLMs' moral and legal reasoning. *Artificial intelligence*, 333: 104145.

Baker, C. L.; Jara-Ettinger, J.; Saxe, R.; and Tenenbaum, J. B. 2017. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1: 1–10.

Gergely, G.; and Csibra, G. 2003. Teleological reasoning in infancy: the naïve theory of rational action. *Trends in cognitive sciences*, 7: 287–292.

Gopnik, A.; and Meltzoff, A. N. 1997. *Words, Thoughts, and theories*. The MIT Press. ISBN 9780262274098.

Guo, X.; Huang, K.; Liu, J.; Fan, W.; Vélez, N.; Wu, Q.; Wang, H.; Griffiths, T. L.; and Wang, M. 2024. Embodied LLM agents learn to cooperate in organized teams.

Hu, J.; Floyd, S.; Jouravlev, O.; Fedorenko, E.; and Gibson, E. 2022. A fine-grained comparison of pragmatic language understanding in humans and language models.

Jara-Ettinger, J.; Gweon, H.; Schulz, L. E.; and Tenenbaum, J. B. 2016. The Naïve Utility Calculus: Computational Principles Underlying Commonsense Psychology. *Trends in cognitive sciences*, 20: 589–604.

Kosinski, M. 2023. Theory of mind might have spontaneously emerged in large language models.

Liu, D.; Wellman, H. M.; Tardif, T.; and Sabbagh, M. A. 2008. Theory of mind development in Chinese children: a meta-analysis of false-belief understanding across cultures and languages. *Developmental psychology*, 44: 523–531.

Martin, A.; and Santos, L. R. 2016. What cognitive representations support primate theory of mind? *Trends in cognitive sciences*, 20: 375–382.

Onishi, K. H.; and Baillargeon, R. 2005. Do 15-month-old infants understand false beliefs? *Science*, 308: 255–258.

Rosati, A. G.; Santos, L. R.; and Hare, B. 2010. *Primate Social Cognition: Thirty Years After Premack and Woodruff*, 117–143. Oxford University Press. ISBN 9780195326598.

Rubio-Fernandez, P.; Berke, M.; and Jara-Ettinger., J. in press. Tracking Minds in Communication.

Shen, S. Z.; Lang, H.; Wang, B.; Kim, Y.; and Sontag, D. 2024. Learning to decode collaboratively with multiple language models.

Ullman, T. 2023. Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks.

Ullman, T.; Baker, C.; Macindoe, O.; Evans, O.; Goodman, N.; and Tenenbaum, J. 2009. Help or hinder: Bayesian models of social goal inference. *Advances in neural information processing systems*, 22.

Wellman, H. M.; and Liu, D. 2004. Scaling of theory-of-mind tasks. *Child development*, 75: 523–541.

Young, L.; Cushman, F.; Hauser, M.; and Saxe, R. 2007. The neural basis of the interaction between theory of mind and moral judgment. 104: 8235–8240.

Yu, C.-L.; and Wellman, H. M. 2024. A meta-analysis of sequences in theory-of-mind understandings: Theory of mind scale findings across different cultural contexts. *Developmental review: DR*, 74: 101162.

Evaluation Paradigms

In our evaluations, we need tractably enumerable paradigms which are capable of eliciting rich ToM inferences. To this end, we develop two logically equivalent domains: *ContainerWorld* and *MovieWorld* as case-studies. Furthermore, we need domains which support parametric prompt construction to map the forward model $\mathcal{F}$ and each mental-state inference $\mathcal{I}$.

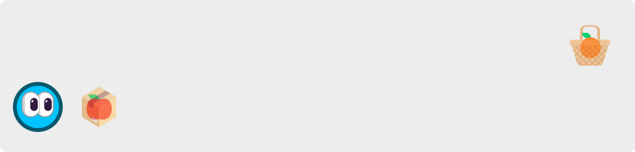

Figure 4: An instance of *ContainerWorld* with an apple in the box and an orange in the basket.

ContainerWorld

ContainerWorld (Fig. 4) has a character that spawns in the south-west corner of a room. This room has a closed box next to the character and a covered basket in the opposite corner, "about 50 fifty steps away". Containers may hold either apples, oranges, or both (apples and oranges) and thus have a state-space $\mathcal{S} = \{\text{apples}, \text{oranges}, \text{apples and oranges}\}$. Due to the partial observability, the character has beliefs about the contents of each container, such that for each container $\mathcal{B} = \{\text{apples}, \text{oranges}, \text{apples and oranges}\}$. Additionally, the character has desires, $\mathcal{D} = \{\text{likes}, \text{dislikes}\}$ towards both apples and oranges. The character may take actions $\mathcal{A} = \{\text{box}, \text{basket}\}$, which entails moving to a container and taking the contents within.

dt= 0
5
10
...
85
90
95

Figure 5: An instance of *MovieWorld* with an action movie in 5 minutes and romance movie in 90 minutes.

MovieWorld

MovieWorld (Fig. 5) has a character at a foreign film festival, thus has difficulty communicating with others. At this film festival, there are two screenings coming up – one starting in 5 minutes and another starting in 90 minutes. Movies screened at this festival are 120 minutes (2 hours) long. The movies' genres are $\mathcal{S} = \{\text{action}, \text{romance}, \text{action-romance}\}$. As a partially observable world, too, the character has beliefs about which movies are playing when such that for each screening $\mathcal{B} = \{\text{action}, \text{romance}, \text{action-romance}\}$. Similarly, the character has desires $\mathcal{D} = \{\text{likes}, \text{dislikes}\}$ towards both action and romance movies. Lastly, the character may take actions $\mathcal{A} = \{05, 90\}$, which entails going to the screening commencing in that amount of time.

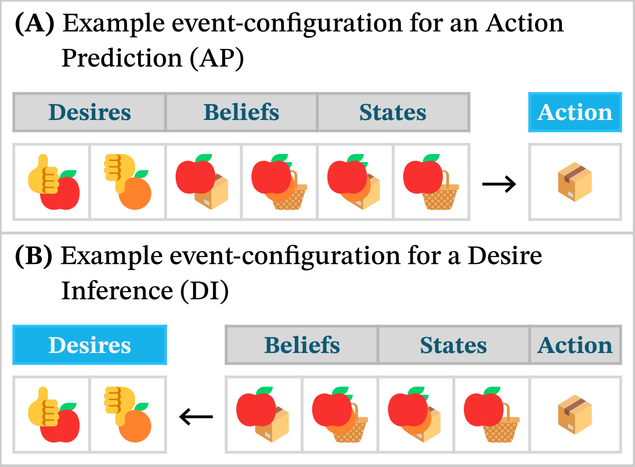
(A) Example event-configuration for an Action Prediction (AP)
Desires
Beliefs
States
Action
→
(B) Example event-configuration for a Desire Inference (DI)
Desires
Beliefs
States
Action
←

Figure 6: A human-like ToM action-prediction and desire inference set in the *ContainerWorld*.

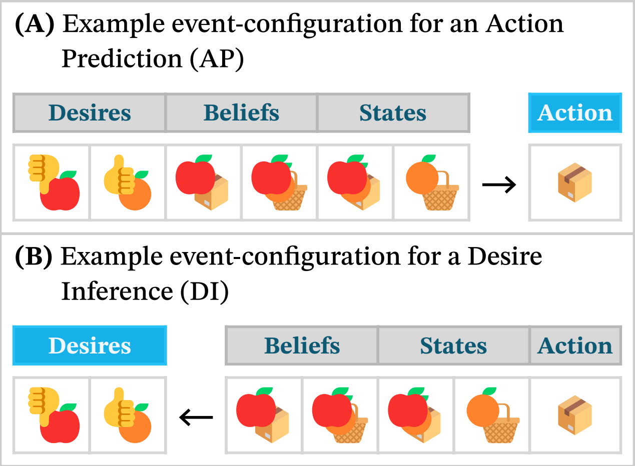
(A) Example event-configuration for an Action Prediction (AP)
Desires
Beliefs
States
Action
→
(B) Example event-configuration for a Desire Inference (DI)
Desires
Beliefs
States
Action
←

Figure 7: A non-human-like-ToM inference set in the *ContainerWorld*. Traditional ToM evaluations would characterize this as a failure in both action prediction and mental-state (desire) inference. Under our approach, this is characterized as a success because the action prediction 7A is coherent with the mental-state inference 7B, conversely, the desire inference coheres with the action prediction.

Large Language Models

We tested this evaluation scheme using GPT-4o via the OpenAI API – using the `openai` package. We used the `gpt-4o-2024-05-13` version, using a temperature of 1, and requesting the top 10 logprobs. Each prediction and inference is a unique query to the LLM. Thus, if the action predictions and desire inferences depicted in Figs. 6 and 7 were given to the LLM, we would issue four queries: two for action prediction with the configuration on the left-hand side (Figs. 6A and 7A) and two for desire inference with the configurations on the right-hand side (Figs. 6B and 7B). Our evaluation approach only works for LLMs which are able to return log-probabilities over next-token predictions, thus LLMs like Claude-3.5 are unable to be evaluated at this time.

LLMs are given a "system prompt" and a "user prompt". The "system prompt" details the goal and response format. While the "user prompt" presents the story describing the world structure and the "belief-state-desire-action" tuple relevant for the prediction or inference.

Mapping the Forward Model $\mathcal{F}$ and Mental-State Inferences $\mathcal{I}$

To map the forward model $\mathcal{F}$ and mental-state inferences $\mathcal{I}$, we translate *ContainerWorld*and *MovieWorld* into prompts. We then query the LLM, GPT-4o in the current work, with each prompt and use the distribution over next tokens as the likelihood of the action prediction and mental-state inference for each task. Below, we describe our prompting setup and how we extract the likelihoods from the next-token distribution.

Parametric Prompt Construction

Each of our paradigms have beliefs $\mathcal{B}$, desires $\mathcal{D}$, states $\mathcal{S}$, and actions $\mathcal{A}$. As illustrated in Fig. 2A, enumerating the forward-model entails creating a permutation of $\mathcal{B}, \mathcal{D}, \mathcal{S}$ to query the LLM. Similarly, as in Fig. 2C, enumerating each mental-state inference entails creating a permutation of $\mathcal{B}, \mathcal{S}, \mathcal{A}$ for desires, $\mathcal{D}, \mathcal{S}, \mathcal{A}$ for beliefs, and $\mathcal{S}, \mathcal{A}$ for belief-desire inferences. Each enumeration consists of two prompts: the "system" prompt and the "user" prompt.

The "system" prompt is depicted in Fig. 8A – the {{ `task` }} is tied to the action prediction and mental-state inference task (Fig. 8B), while {{ `schema` }} is derived from the task and paradigm. The {{ `schema` }} is a JSON-serialization mapping each input to the unique constituents of $\mathcal{B}, \mathcal{D}, \mathcal{S}, \mathcal{A}$. In Fig. 8C, we show the schema which informs an LLM what to respond with for desire inferences in *ContainerWorld*. Similarly, in Fig. 8D, we should the schema for belief inferences in *MovieWorld*. We use schemas like these for each prediction and inference task to maximize response structure and minimize extraneous text-generation.

The "user" prompt (Fig. 9A) is a concatenation of a "context" (background details framing the paradigm), the prediction or inference inputs, and a "query" for the paradigm output(s). The "query" is both task- and paradigm-dependent, while "context" and the other inputs only vary across paradigms. The prompt components for *ContainerWorld* and *MovieWorld* are detailed in Fig. 9B and Fig. 9C, respectively. We provide a full example prompt construction in Figs. 10 and 11.

(A) “System” Prompt given to LLM

`{{ schema }}` is both task- and paradigm-dependent.

```
Your task is to tell us {{ task }} using the JSON
Schema provided below.

{{ schema }}
```

(B) Text Replacements for `{{ task }}` in (A)

Task	Content
AP	the action someone you’re observing would take
DI	the desires of someone you’re observing
BI	the beliefs of someone you’re observing
BDI	the beliefs and desires of someone you’re observing

(C) Abbreviated `{{ schema }}` for **DI** in *ContainerWorld*

```
{
  "desires": {
    "apples": {
      "type": "string",
      "enum": [
        "likes",
        "dislikes"
      ]
    },
    "oranges": {
      "type": "string",
      "enum": [
        "likes",
        "dislikes"
      ]
    }
  }
}
```

(D) Abbreviated `{{ schema }}` for **BI** in *MovieWorld*

```
{
  "beliefs": {
    "screening05": {
      "type": "string",
      "enum": [
        "action",
        "romance"
      ]
    },
    "screening90": {
      "type": "string",
      "enum": [
        "action",
        "romance"
      ]
    }
  }
}
```

Figure 8: System Prompt: In 8A, we illustrate the prompt template used across *ContainerWorld* and *MovieWorld*. {{ `task` }} is replaced with the content shown in 8B, based on the current task. In 8C and 8D, we illustrate example schemas used for desire and belief inferences.

Mapped Distribution Construction

While LLMs are able to assign a likelihood to their predictions by text, a more direct measure is to use the likelihood from the next-token distribution. We do this by searching for the target of a given task and extracting their log-probabilities.

Consider the example depicted in Fig. 10 – a joint belief-desire inference set in *ContainerWorld*. The LLM’s response will be JSON according with the schema shown in Fig. 10A. We then search for the responses generated for each of `desires.apples`, `desires.oranges`, `beliefs.box`, and `beliefs.basket`. Once found, we extract the token log-probabilities for the eligible values (denoted in the `enum` field of the schema) and normalize their scores to create a distribution over eligible values, which becomes the likelihood function for each component. We note that our current implementation overestimates the likelihood of words which are tokenized into more than one token (e.g., “dislikes” tokenizes to [“dis”, “likes”]). However, we do not believe this substantively impacts our findings as this shortcoming applies across all tasks and paradigms. Furthermore, we find no substantive correlation between almost all action prediction Bayesian inversions and mental-state inferences (notably, DI in *MovieWorld* is moderately correlated).

The distributions retrieved in this way are used for both evaluation measures: (1) if the mental-state inference produces the observed action is considered “valid” (the “validity” evaluation described earlier) and (2) the correlation of the direct mental-state inferences with the Bayesian inversion of the forward model $\mathcal{F}$ (the “Bayesian” evaluation described earlier).

(A) "User" Prompt given to LLM

Each component below is paradigm-dependent. query is both task- and paradigm-dependent.

AP: context ⧺ beliefs ⧺ desires ⧺ states ⧺ query

DI: context ⧺ beliefs ⧺ states ⧺ actions ⧺ query

BI: context ⧺ desires ⧺ states ⧺ actions ⧺ query

BDI: context ⧺ states ⧺ actions ⧺ query

(B) ***ContainerWorld***

Component	Content
context	John is standing in the corner of a large room.
beliefs	John believes that the closed box has $\mathcal{B}$ and that the covered basket has $\mathcal{B}$.
desires	John $\mathcal{D}$ apples and he $\mathcal{D}$ oranges.
states	Within arm's reach, there is a closed box. The closed box is filled with $\mathcal{S}$. In the opposite corner, about fifty steps away, there is a covered basket. The covered basket is filled with $\mathcal{S}$.
actions	John wants to eat fruit, so he goes to the $\mathcal{A}$ and takes the fruit inside.
query	**AP:** He wants to eat a single piece of fruit. Which container would John take fruit from?
query	**DI:** What are John's desires?
query	**BI:** What are John's beliefs?
query	**BDI:** What are John's beliefs and desires?

(C) ***MovieWorld***

Component	Content
context	Alex is at a foreign film festival. He has a surface-level understanding of the local language, but doesn't know enough to be conversational. This film festival only screens movies which are 120 minutes long.
beliefs	Alex believes that the screening starting in 5 minutes is $\mathcal{B}$ movie and that the screening starting in 90 minutes is $\mathcal{B}$ movie.
desires	Alex $\mathcal{D}$ action movies and he $\mathcal{D}$ romance movies.
states	There is $\mathcal{S}$ movie starting in 5 minutes and $\mathcal{S}$ movie starting in 90 minutes.
actions	Alex wants to watch a screening, so he goes to the screening starting in $\mathcal{A}$ minutes.
query	**AP:** He wants to watch a movie. Which screening should Alex go to?
query	**DI:** What are Alex's desires?
query	**BI:** What are Alex's beliefs?
query	**BDI:** What are Alex's beliefs and desires?

Figure 9: User Prompt: In 9A, we show the component-definition of the user prompt template. For action-prediction ($\mathcal{F}$), we would use the "context", "beliefs", "desires", "states", and the "query" for **AP** (action prediction). These values are then concatenated to form a prompt which an LLM is then expected to respond to according to the `{{ schema }}` specified in the system prompt (Fig. 8A). We detail the specifics for *ContainerWorld* in 9B and for *MovieWorld* in 9C.

Example prompt pair for *ContainerWorld* on **BDI**

We color variadic replacements based on their component

(A) "System" Prompt

Your task is to tell us the beliefs and desires of someone using the JSON Schema provided below.

```
{
  "desires": {
    "apples": {
      "type": "string",
      "enum": [
        "likes",
        "dislikes"
      ]
    },
    "oranges": {
      "type": "string",
      "enum": [
        "likes",
        "dislikes"
      ]
    }
  },
  "beliefs": {
    "box": {
      "type": "string",
      "enum": [
        "apples",
        "oranges"
      ]
    },
    "basket": {
      "type": "string",
      "enum": [
        "apples",
        "oranges"
      ]
    }
  }
}
```

(B) "User" Prompt

John is standing in the corner of a large room. Within arm's reach, there is a closed box. The closed box is filled with apples. In the opposite corner, about fifty steps away, there is a covered basket. The covered basket is filled with apples. What are John's beliefs and desires?

Figure 10: An example prompt pairing for *ContainerWorld* on the belief-desire inference (BDI) task.

Example prompt pair for *MovieWorld* on **BDI**

We color variadic replacements based on their component

(A) "System" Prompt

Your task is to tell us the beliefs and desires of someone using the JSON Schema provided below.

```
{
  "desires": {
    "action": {
      "type": "string",
      "enum": [
        "likes",
        "dislikes"
      ]
    },
    "romance": {
      "type": "string",
      "enum": [
        "likes",
        "dislikes"
      ]
    }
  },
  "beliefs": {
    "screening05": {
      "type": "string",
      "enum": [
        "action",
        "romance"
      ]
    },
    "screening90": {
      "type": "string",
      "enum": [
        "action",
        "romance"
      ]
    }
  }
}
```

(B) "User" Prompt

Alex is at a foreign film festival. He has a surface-level understanding of the local language, but doesn't know enough to be conversational. This film festival only screens movies which are 120 minutes long. There is a romance movie starting in 5 minutes and an action-romance movie starting in 90 minutes. What are Alex's beliefs and desires?

Figure 11: An example prompt pairing for *MovieWorld* on the belief-desire inference (BDI) task.