The Neuro-Symbolic Inverse Planning Engine (NIPE): Modeling Probabilistic Social Inferences from Linguistic Inputs

Lance Ying [1] [2] **Katherine M. Collins** [3] **Megan Wei** [1] **Cedegao E. Zhang** [1] **Tan Zhi-Xuan** [1]
Adrian Weller [3] **Joshua B. Tenenbaum** [1] **Lionel Wong** [1]

Abstract

Human beings are social creatures. We routinely reason about other agents, and a crucial component of this social reasoning is inferring people's goals as we learn about their actions. In many settings, we can perform intuitive but reliable goal inference from language descriptions of agents, actions, and the background environments. In this paper, we study this process of language driving and influencing social reasoning in a probabilistic goal inference domain. We propose a neuro-symbolic model that carries out goal inference from linguistic inputs of agent scenarios. The "neuro" part is a large language model (LLM) that translates language descriptions to code representations, and the "symbolic" part is a Bayesian inverse planning engine. To test our model, we design and run a human experiment on a linguistic goal inference task. Our model closely matches human response patterns and better predicts human judgements than using an LLM alone.

1. Introduction

Humans are deeply social creatures, finely attuned to the inner lives of other people (Wellman, 1992; Saxe, 2006; Tomasello, 2010). We not only think, but think routinely about each other. We imagine what someone is trying to accomplish and why are trying to accomplish it; we try to parse what someone really believes and whether they might be mistaken; and we anticipate people's ultimate goals and upcoming actions, stepping in to assist or sometimes even to try to hinder them. This capacity for *social reasoning*, broadly, is *abstract* (we can observe someone's actions in the world, but cannot directly see what they think or want) and relies on a fine grained knowledge of distinct but highly *interrelated* mental states within others (we know that someone's beliefs, desires, and goals relative to the world itself all collectively influence what they might choose to do.)

This paper considers how **language can inform social reasoning**, with a particular focus on **understanding language about people's actions, plans, and goals**. Language is an especially powerful means of conveying the abstract, inter-related nature of concepts about agents in the world—we can talk about someone's actions (*I saw Annie heading into the living room*), directly convey their mental states (*Annie desperately wants a toy in there*), describe relevant but abstract aspects of the world in which they act (*that cupboard Annie is trying to open is locked inside*), or even pose the questions we want others to reason about (*which toy do you think she wants the most?*). Language about any one of these aspects informs how we might reason about someone else in holistic ways, changing our understanding and downstream predictions about what they might know, want, or do.

Empirical evidence suggests that we think intuitively about each other as largely *rational, goal-directed intelligent agents* (Dennett, 1981; Csibra et al., 1999; Baker et al., 2009; Baillargeon et al., 2016). Within computational cognitive science and AI, a considerable body of prior work formalizes these intuitions by modeling agents as *rational actors and planners* who choose actions to achieve specified goals or optimize rewards, and in turn models many inferences about mental states as *inverse planning* from observations of actions back to their latent goals, utility functions, or beliefs about world states (Jara-Ettinger et al., 2016; Baker et al., 2017; Zhi-Xuan et al., 2020; Ying et al., 2023). Despite considerable successes in predicting distinct human inferences and modeling cooperative behavior in human-computer interactions (Carroll et al., 2019; Mozannar & Sontag, 2020; Alanqary et al., 2021; Bansal et al., 2021; Wilder et al., 2021; Shneiderman, 2022), these approaches—largely *non-linguistic* models that draw inferences from images or demonstrations of agent actions—face scalability and specification challenges inherent to the ab-

---

[1]Massachusetts Institute of Technology, Cambridge, USA [2]Harvard University, Cambridge, USA [3]University of Cambridge, Cambridge, UK. Correspondence to: Lance Ying <lanceying@seas.harvard.edu>.

Linguistic Goal Inference Scenarios

World Configuration

There are 3 trophies, gold, silver and bronze. The bronze trophy is located behind a red door.

Alice is located 3 steps north from the red door, 4 steps east from the green key.

Abstract World Dynamics

Each door can be unlocked by a key with the corresponding color.

Keys can only be used once

Agent Actions

You observe Alice walking two step South

You observe Alice picking up 2 keys

Goal Inference Query

Which trophy is Alice's goal?

LLM Language to Code

Symbolic Generative Model Representation

Generative Model over spatial layout

xloc(Alice), yloc(Alice) ~ uniform([1:col], [1:row])
xloc(green), yloc(green) = xloc(Alice) + 4, yloc(Alice)
xloc(red), yloc(red) = xloc(Alice), yloc(Alice) + 3

PDDL problem definition

```
(define (problem game_show1)
(:domain game_show)
(:objects Alice - agent key1- key yellow - colors
  gold silver bronze - trophy)
(:init (locked door1)
  (= (iscolor door1) red)
  (= (iscolor key1) red)
```

Domain Definition

```
(:action unlock
:parameters (?a - agent ?k - key ?d - door)
:precondition (and (has ?a ?k) (locked ?d)
  (exists (?c - color) (and (iscolor ?k ?c) (iscolor ?d ?c)))
:effect (and (not (has ?a ?k)) (not (locked ?d)))
```

Action Simulation

["(South Alice)", "(South Alice)"]

(exists (?k1 ?k2 - key) (and (has Alice ?k1) (has Alice ?k2) (not (= ?k1 ?k2))))

?trophy | plan(Alice)

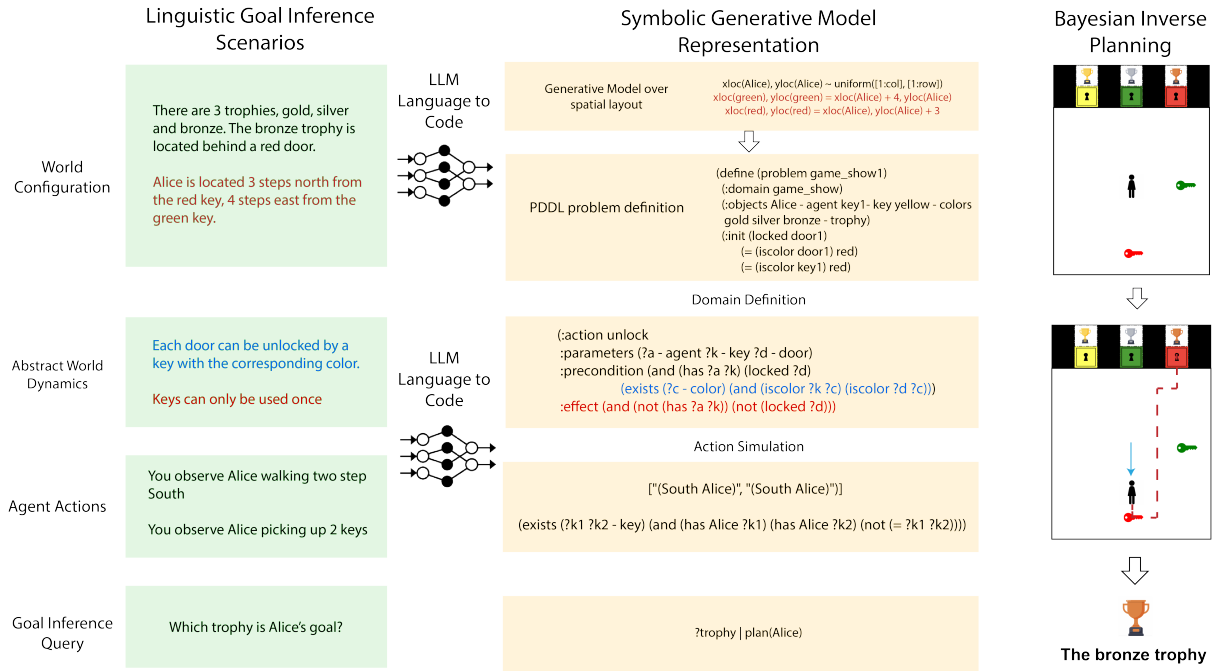## Bayesian Inverse Planning

The bronze trophy

Figure 1. Overview of our model, the *Neuro-Symbolic Inverse Planning Engine* (NIPE), which reasons about linguistic goal inference scenarios. (*Left, green*) We design and measure human judgments on a *linguistic goal inference* domain (inspired by (Zhi-Xuan et al., 2020), which asks about the final goal of a participant on a complex "gameshow" obstacle course involving trophies behind locked doors. Inferring goals requires reasoning jointly about language that describes the overall world configuration, specific world dynamics, and the agent's actions. Our model works by first using a large language model to *translate* language into symbolic program semantics that collectively condition a probabilistic generative model over possible environments and agents (*Middle, yellow*), and answers queries via *Bayesian inverse planning* (*Right*) in the conditioned probabilistic model, inferring likely goals using a nested planner to model the agent.

stract, complex, and interrelated nature of social reasoning tasks. Modeling agents as goal-directed planners requires implicitly or explicitly representing the *world model* they are planning over, which is often precisely defined a priori (as in Baker et al. (2017) and Zhi-Xuan et al. (2020)) or learned from expensive exploration (as in Lian et al. (2021)). Inferring mental states, or even posing the task itself, is challenging when framed as an inductive problem from images and demonstrations alone (as in Shu et al. (2021)).

In part due to these challenges, the recent advances made by *large language models* (LLMs) have driven an explosion of interest in how these models can be used to plan (Valmeekam et al., 2022) or draw social inferences (Bubeck et al., 2023; Shapira et al., 2023) from language—they suggest a new frontier of cognitive models and AI systems that can leverage language to specify abstract world knowledge, domain-specific world dynamics, and direct information about agents and their mental states. However, early efforts to apply LLMs directly to social reasoning problems often take a highly simplified approach to evaluating their capacity for social reasoning or their *"theory of mind"*, drawing broad conclusions based on a narrow set of templated cognitive tasks (like the famous "Sally Anne tasks"). Other eval-

uations caution against drawing general conclusions from such restricted task settings, pointing out that modifications and extensions beyond these simple problem settings can produce counterintuitive results that contradict the holism and generality of these inferences (Shapira et al., 2023; Ullman, 2023); and several recent evaluations suggest that robust planning in general remains an outstanding challenge for large language models (Valmeekam et al., 2022; Bubeck et al., 2023). Several recent approaches seek to address these challenges by augmenting LLMs with interfaces to external modules for representing and simulating complex aspects of world state, particularly using classical AI representations and engines for forward planning (Collins et al., 2022; Liu et al., 2023a; Lyu et al., 2023). Emulating human-like social reasoning and knowledge, however, requires models that can robustly draw *unified inferences about actions, mental states, and the world* (Wong et al., 2023), relating plans and beliefs in holistic, flexible, and ideally interpretable and safely deployable way to language; and linguistic evaluations that move in turn towards the richness and breadth of human-like social reasoning, flexible and extensible enough to incorporate these interrelated aspects of human mental states and world knowledge.

We focus specifically in this paper on **understanding language that describes inverse planning and goal inference problems**—we consider how language can convey information about the world itself, and the actions within it, in order to drive inferences about the unseen goals motivating these actions. We make the following contributions:

1. We introduce a new model, the *Neuro-symbolic Inverse Planning Engine* (NIPE) that leverages LLMs to *translate language into program expressions* that compositionally construct and condition a *probabilistic generative model over agents and environments*, then uses *Bayesian inverse planning* to perform goal inference over this symbolic representation, expanding upon the paradigm of *rational meaning construction* for language-informed cognition (Wong et al., 2023).

2. To evaluate this model, we compare our model against *human social reasoning on a linguistic inverse planning domain* inspired by prior cognitive work in social cognition (Zhi-Xuan et al., 2020). We design this domain to require jointly reasoning about agent actions and varying information about the world itself.

3. We find that our model *well-predicts human judgments* across a variety of distinct social reasoning scenarios. We also find that our model *outperforms LLM baselines used to directly perform the same goal inference tasks*. We find that the largest LLM alone (GPT-4 (OpenAI, 2023)) correlates moderately well with certain human judgments on this domain, particularly on simpler scenarios. We discuss trends between LLMs of different sizes and on different task variants in our stimuli.

2. Linguistic inverse planning tasks and human experiment

We begin by describing the domain of tasks and human experiment we use throughout this paper, to provide intuition for the modeling experiments that follow. We construct a domain of **linguistic inverse planning stimuli** designed to evaluate how humans and models **infer the goals of agents acting within varied environments**. We design stimuli to convey varying information about the *world configuration and dynamics* in which agents behave, as well as the *observed actions of agents* with unknown goals.

Our tasks are inspired by the spatial navigation stimuli used in prior social reasoning experiments (Zhi-Xuan et al., 2020; Alanqary et al., 2021). Our experiments adapt this basic domain into a set of linguistic inverse planning stimuli, and introduce a range of variations designed to test distinct world dynamics, spatial information, and vague or incomplete information about agent actions conveyed linguistically.

2.1. Planning and inverse planning concepts in the Gameshow Domain

Each stimuli in our experiment (Table 1) begins with a description of a "*gameshow*" scenario in which a player is trying to reach one of several goal *trophies* placed around an obstacle course. This obstacle course also includes *locked doors* that can only be unlocked with keys placed around the course, and which additionally require multi-step reasoning about an agent's longer term plans. Based on each scenario description, which includes varying information about the particular obstacle course and the actions of an agent within it, we pose a linguistic goal inference query: *which of the possible trophies do you think the player trying to reach?*

We design each stimulus to span several core conceptual categories relevant to reasoning about agents and their actions in the world (Table 1, *top*), including:

- **World configuration**: information about goal trophies, keys, and their attributes, as well as their relative locations, at differing levels of specificity.
- **Abstract world dynamics**: causal dynamics within the environment, such as rules determining which keys can open which doors.
- **Agent actions**: descriptions of an agent's behavior before they reach the goal, like *picking up keys* or *walking* in different directions towards keys or trophies.

2.2. Task variants

We design four distinct variants of gameshow stimuli for a total of *n=18* stimuli, which vary the complexity of the specified world configuration, the world dynamics, and the type of observed agent actions (excerpts from stimuli in each variant are shown in Table 1, *bottom*):

- **Generic keys:** Goal trophies are placed behind locked doors, and must be unlocked by keys placed around the obstacle course. Any key on the course can be used to open any other door.
- **Same-colored keys:** This variant adds additional complexity in the world dynamics. Keys and doors have colors. Doors can only be unlocked by keys that are the same color as the door.
- **Different-colored keys:** In this variant, doors can only be unlocked by keys that are a *different* color from the door.
- **Spatial:** This variant contains specific spatial descriptions of the obstacle course and agent actions. Due to the added spatial complexity, in these environments, there are no keys or doors; agents can walk freely towards various goal trophies.

Linguistic planning and inverse planning concepts across tasks	
World configuration	There are three trophies placed through this obstacle course: gold, silver, and bronze; The gold trophy is currently 5 steps East of Alice; The silver trophy is behind 1 door.
World dynamics	On this course, each lock can only be opened by a key that is the same color as that lock; In this game, players lose points if they pick up more keys than needed.
Agent actions	You see Alice picking up a green key; You see Alice walking a step towards the South; In this game, players lose points if they pick up more keys than needed.
Goal inference	Considering each of the possible trophies on this course, how likely is it that Alice is trying to reach this trophy?
Task variations	
Generic Keys	There are three trophies placed throughout this obstacle course: gold, silver, or bronze. To get to the gold trophy, you need to unlock 2 different doors. To get to the silver trophy, you need to unlock 1 door. The bronze trophy isn't behind a door. Keys to unlock the doors are placed around the obstacle course. **All of the keys on this course work on all of the locks.** You see Alice pick up 1 key.
Same-Colored Keys	There are three trophies placed throughout this obstacle course: gold, silver, or bronze. The gold and silver trophies are both located behind a yellow door. To go to the bronze trophy, you need to go through a red door. On this course, **each lock can only be opened by a key that is the same color as that lock**. You see Alice pick up a red key.
Different-Colored Keys	There are three trophies placed throughout this obstacle course: gold, silver, or bronze. The gold and silver trophies are both located behind a green door. To go to the bronze trophy, you need to go through a red door. On this course, **each door can be unlocked by any key that is a different color than the door**. You see Alice pick up a red key.
Spatial	There are three trophies placed throughout this obstacle course: gold, silver, or bronze. Alice is a participant on this course. The gold trophy is currently **5 steps East of Alice**. The silver trophy is **3 steps South from her**. The bronze trophy is **5 steps further West of the silver trophy**. You see Alice walking a step towards the South.

Table 1. Example language from the linguistic "gameshow" goal inference domain. (*Top*) Each scenario contains language that spans a variety of interrelated concepts relevant to the final goal-inference query, ranging from information about the particular environment configuration to varying information about an agent's actions. (*Bottom*) We design four distinct task variants that vary in complexity and type of the environment description and the spatial details of the agent's actions.

2.3. Human experiment

We collect and evaluate human judgments on these goal inference tasks. We recruit a total of n=46 participants from Prolific [1]; each viewed a random batch of 7 stimuli from a specific task variant, and for each stimulus, produced goal inference judgements about each possible goal trophy on a 1-7 Likert scale (*1: definitely not this trophy; 7: definitely this trophy*). Scenarios were presented for a fixed amount of time before an answer was permitted (90 seconds in the *Spatial* condition; 45 seconds in the others). Each scenario was seen by between 14-16 participants. Participants were native English speakers from the USA/UK.

[1]We exclude 6 participants as less than 30% of their MAP results agreed with consensus; i.e. they disagree with the majority on which goal is the most likely for at least 5 out of 7 stimuli.

3. Our Model: The Neuro-Symbolic Inverse Planning Engine (NIPE)

We now introduce our model, the *Neuro-Symbolic Inverse Planning Engine* (NIPE). While we focus on goal inference tasks, we describe a general framework for conditioning on the diverse range of linguistic information in these stimuli—varying information about the spatial environment, world dynamics, and agent actions—that supports inferences about agents based on a principled, probabilistic framework for Bayesian inverse planning.

Our core contribution is to show how we can approach complex *linguistic* scenarios about agents by leveraging structured, symbolic representations and inference methods (modeling agents as explicit *planners*, and performing principled Bayesian inference over complex models) from the existing cognitive and AI social reasoning literature (eg. (Baker et al., 2017; Zhi-Xuan et al., 2020)). We do so by

translating from language into program expressions that condition a **symbolic generative model over agents and environments**, instantiating the paradigm of *rational meaning construction* for language-guided cognition (Wong et al., 2023). This approach also addresses an important open challenge for prior non-linguistic (inverse) planning models: We can condition on abstract information about the world that is specified in language, such as information about *where objects are located* or *which keys open which doors*.

Our model (shown in schematic form in Figure 1) modularly decomposes the problem of goal inference from language into three general subproblems: *language-to-code translation*, which we instantiate generally by using a LLM as a few-shot semantic parser (similar to (Zhang et al., 2023; Liu et al., 2023b; Collins et al., 2022; Wong et al., 2023)); the construction of a *probabilistic generative model over agents and environments* from language; and *goal inference through Bayesian Inverse Planning* in this probabilistic model (using an nested planning and inference algorithm like that in (Zhi-Xuan et al., 2020)). We now describe each problem in detail.

3.1. Language-to-code translation using LLMs

Following Wong et al. (2023), the crux of our approach is to show how the different kinds of language in our scenarios can each be translated into program expressions that collectively construct and condition a holistic probabilistic model. We begin, however, by briefly describing the language-to-code translation itself, which we implement by using LLMs to *translate sentences into code*, which compositionally construct a symbolic model over an extended, multi-sentence linguistic scenario (see Figure 1, *LLM Language to Code*).

Our translation approach leverages the few-shot prompting capabilities of LLMs trained jointly on language and code. We choose *GPT-3.5* (Ouyang et al., 2022) specifically to show that a *smaller* LLM can be used to implement this more restricted task, rather than using the LLM itself to internally represent and reason over these scenarios. Future work should explore whether even smaller and more restricted language-to-code models can be used. Our approach builds on the methods used in Zhang et al. (2023), which uses an LLM to translate language into probabilistic programs, along with other work showing that LLMs can translate language into code when given paired few-shot examples (Liu et al., 2023a; Wong et al., 2023). For each stimulus, we construct a prompt containing 2-3 randomly sampled examples of paired language and code from held out examples on our domain. Models were sampled with a temperature of 1.2. To ensure robustness, we rejection sample translations to ensure that they pass a number of syntactic and semantic tests (e.g., to ensure that translated code is valid and executable.)

3.2. Constructing a probabilistic generative model over agents and environments from language

Now we describe how the different kinds of information relevant to this domain can be each translated into different kinds of code within a unifying generative model. Following Zhi-Xuan et al. (2020), we model information about the world dynamics, such as action preconditions and effects, using the Planning Domain Definition Language (PDDL), a standardized representation language widely used in robotics and automated planning (Aeronautiques et al., 1998), and which can b extended with modular semantics to efficiently represent spatial domains such as ours (Gregory et al., 2012; Zhi-Xuan, 2022).

While prior work has considered how different subproblems related to planning can be extracted via language-to-code (Collins et al., 2022; Liu et al., 2023b; Silver et al., 2023), our aim here is to translate language not only to distributions over possible worlds — i.e. possible spatial layouts and domain rules — but also to inferences over an agent's actions. We describe each aspect of translation below:

Conditions on spatial environments First, we prompt LLM to translate sentences like "The bronze chest is located behind a red door." into a `CONDITION` statement that describes this arrangement. We then sample concrete grid-world maps from a generative model over maps that satisfy this.

To ensure that maps are valid for our environment, we define and condition a relatively restricted generative model over possible maps and trophy placements that ensures that these basic validity conditions are met. Each sampled map from this generative model is a concrete PDDL scene representation that describes the locations of the initial agent, objects, and their attributes.

Conditions on domain rules Now, consider sentences like "Each door can be unlocked by a key with the corresponding color." These can be straightforwardly translated into operator definitions expressed in PDDL (see Figure 1), which define an action and its necessary parameters, preconditions for trying this action, and effects on the environment.

Collectively, the scene representation and the domain rules define a complete, concrete planning domain defined in PDDL. As in Zhi-Xuan et al. (2020), this domain model is nested within a larger Bayesian model over agents acting in this environment. The benefit of using PDDL is that we can easily model the agents themselves using classical planning algorithms over the PDDL domain representation and we can easily transfer this model to a different problem setup by modifying the operator definitions or simply choosing a different pre-defined PDDL domain.

Conditioning a generative model over agent subgoals and actions from language Finally, we consider how to translate sentences that describe agents' actions into conditions on a model over agents. We distinguish between language like "You observe Alice picking up two keys" which is treated as a condition on subgoals, and "You observe Alice walking 3 steps East.", which is treated as a low-level action sequence.

3.3. Goal Inference through Bayesian Inverse Planning

Once we have an initial PDDL environment representation s_0 and a sequence of agent actions $a_{1:t}$, we can perform Bayesian inverse planning to infer a posterior distribution over goals g given actions $a_{1:t}$ and the initial state s, which is proportional to the joint distribution over goals and actions:

$$P(g|s_0, a_{1:t}) \propto P(g, a_{1:t}|s_0) = P(g|s_0)P(a_{1:t}|g)$$

Here $P(g|s_0)$ defines the prior distribution over possible goals in the initial state s_0, and $P(a_{1:t}|g)$ is a model of how the agent selects actions $a_{1:t}$ given a particular goal g. For the goal prior, we model agents as being more likely to pick up goals that are closer to them, as determined by the cost $C(g, s_0)$ of the shortest plan to each goal:

$$P(g|s_0) \propto \frac{1}{C(g)}$$

To model actions, we assume $P(a_{1:t}|g)$ decomposes into the probability of selecting an action a_t at each state s_t, given by a Boltzmmann-rational policy π:

$$P(a_{1:t}|g) = \prod_{\tau=1}^{t} \pi(a_\tau|s_t, g)$$
$$\pi(a_t|s_t, g) = \frac{\exp \beta Q_g(s_t, a_t)}{\sum_{a'_t} \exp \beta Q_g(s_t, a'_t)}$$

where $Q_g(s_t, a_t)$ denotes the (negative) cost of the reaching the goal g by taking action a_t from s_t, and β is a rationality parameter, also called the inverse temperature. This policy models agents as approximately rational in how they act, since actions which reach the goal more quickly (higher Q_g) are more probable, with higher values of β leading to more optimal actions. Although this model does not account for forms of approximate rationality (Evans et al., 2016; Shah et al., 2019; Alanqary et al., 2021), it is a widely used choice in (Bayesian) inverse planning and inverse reinforcement learning that can account for non-systematic deviations from optimality (Ramachandran & Amir, 2007; Ziebart et al., 2008; Baker et al., 2009). To compute Q_g for each state s_t and action a_t encountered during inference, we use A* search as an optimal-cost planner in the PDDL domain to determine the cost of the optimal plan.

With these components of the model specified, goal inference can be performed *exactly* by computing the joint probability $P(g, a_{1:t}|s_0)$ of each possible goal g with the observed sequence of actions $a_{1:t}$, which can then be normalized to obtain the posterior distribution $P(g|s_0, a_{1:t})$.

4. Model Experiments

We compare our model with human judgements on the gameshow domain of social reasoning tasks. We also evaluate two LLM baselines to assess the value of an explicit Bayesian inverse planning model.

For our model, we simulate answers to each goal inference stimulus (*which trophy will Alice pick up?*), outputting a *posterior probability* over all of the possible trophies (probabilities sum to 1 over all trophies). For each stimulus, we translate the linguistic scenario description of the world configuration, world dynamics, and agent behavior into symbolic conditions on the generative model (by sampling from the LLM and rejecting invalid code), and perform goal inference with the generated code to calculate the posterior over the trophies. Since goal inference can be performed without sampling, and since the space of maps consistent with a linguistic description is highly restricted, we found in pilot experiments that our model showed limited variability across LLM and generative model samples. Therefore, for each stimulus, we computed our results using a single valid sample of PDDL domain rules and spatial environments.

We also conduct two baseline experiments using two different LLMs (`gpt-3.5-turbo-0301`, the dialogue-tuned variant of GPT-3.5 (Ouyang et al., 2022), and `gpt-4-0314`, the March snapshot of GPT-4 (OpenAI, 2023)) to directly provide goal inference judgments for each stimulus. We assess LLMs using exactly the same conditions as the human experiment, using a zero-shot prompt where the LLM is given the full linguistic experimental set up (including instructions explaining the task) shown to human participants, concatenated with the scene scenario and query for each stimulus, and a short prompt specifying the answer format. To account for answer variability, we sample *n = 30* answers for each stimulus, rejecting any answers that are not in the correct format. Exact prompt formats can be found in the Appendix.

5. Results and discussion

We evaluate our model and baselines in comparison to human judgments, both across all of the stimuli (Figure 2) and within each task variant category (Figure 3). We find that:

Our model (NIPE) correlates strongly overall with human judgments across the goal inference experiment. We calculate correlations between the *posterior probability judgments* produced by our model and the mean per stimulus judgments (Likert scale ratings) across human subjects, and find that our model is strongly correlated with human judgments in the experiment overall (Figure 2, *Our Model*, R=0.927, 95% CI=(0.877, 0.953)). All 95% CI are bootstrap sampled confidence intervals each with 1000 samples.

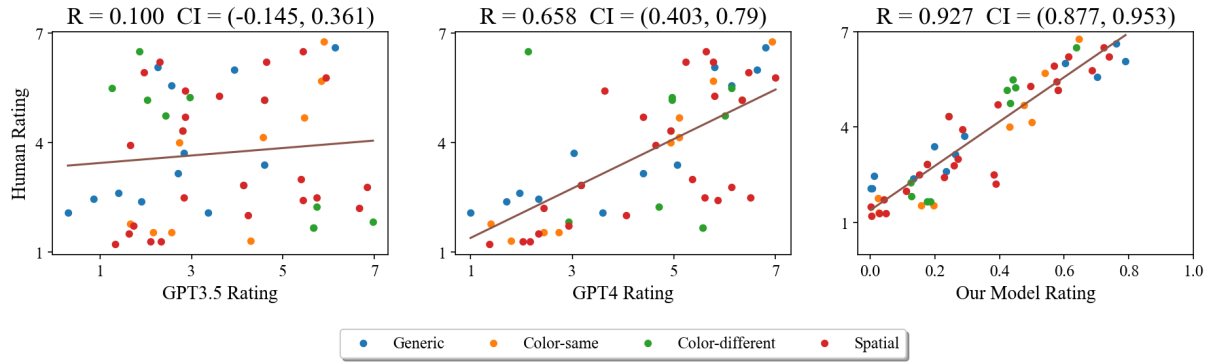*Figure 2.* Correlation between NIPE and baseline LLMs vs. human judgments, aggregated across the goal-inference stimuli. Our model (*Right*) produces posterior probabilities over goals that correlate strongly overall with human judgments; GPT-4 correlates moderately overall (*Middle*) and GPT-3.5 (*Left*) correlates poorly across the tasks. Color-coded points show stimuli from specific task variants.

The largest LLM baseline (GPT-4) correlates moderately with human judgments overall; GPT3.5 correlates poorly. We additionally calculate correlations between LLMs prompted directly to produce goal inference judgments (mean over sampled LLM judgements for each stimulus, and the mean over human judgments for each stimulus across participants). In comparison to our model, we find that GPT-4 correlates well, but more weakly, with human judgments (Figure 2, *GPT-4 Rating*, R=0.658, CI=(0.403, 0.79)) overall, and GPT-3.5 correlates much more poorly with human judgments overall (Figure 2, *GPT-3.5 Rating*, R=0.100, CI=(-0.145, 0.361)). These aggregate trends are better understood by examining correlations between human judgments within each task variant, as we discuss next.

NIPE correlates strongly with human judgements *within* each task variant; GPT-3.5 and GPT-4 are less robust to complex world dynamics and spatial information. Inspecting trends *within* each individual task variant (Figure 3) provides insight into when models deviate from humans.

NIPE consistently correlates with human judgments across *all* task variants, from the simpler *Generic* stimuli to the more complex *Colored* and *Spatial* stimuli (Figure 3, bottom). In contrast, on the more complex tasks—particularly the *Colored-Different* variant, which specifies a new, counterintuitive rule particular to this domain (*doors can only be opened by keys that are a **different** color from the door*), and the *Spatial* variant, which requires considering how an efficient, path-planning agent would navigate towards their goal—the LLM baselines show revealingly different performance in comparison to human reasoning.

As seen in Figure 3, GPT-3.5 actually correlates *negatively* with human judgments on *Colored-Different*—it both ignores the stated rule about the different colored keys and appears to make the opposite, incorrect assumption, perhaps based on simpler color-based pattern matching; and

shows nearly zero correlation with human judgments on the explicit *Spatial* path-planning variant. In comparison, GPT-4 correlates well with human judgments on both of the simpler task variants, and shows a moderate (but positive) correlation on the *Spatial* variant (though we discuss stimuli-specific failures next, suggesting that GPT-4 may still be using simpler distance heuristics rather than robustly modeling an efficient path-planning agent.) On the *Colored-Different* domain, GPT-4 has nearly zero-correlation, suggesting that it is highly uneven in taking the specific rule about key colors into account—as shown in Figure 3, it predicts human judgments well in some cases, but produces nearly the opposite overall judgment in others, suggesting it has ignored the rule.

Humans make fine-grained assumptions about agents as rational planners; LLMs do not always do so. Our model explicitly reasons about agents as rational, goal-directed actors and planners. Do LLMs do so as well? Qualitative analysis of individual stimuli highlights subtle but important discrepancies between human assumptions and LLM-based goal inferences on the most complex *Spatial* stimuli. We highlight a representative example here:

> *There are three trophies placed throughout this obstacle course: gold, silver, or bronze. Alice is a participant on this course. The gold trophy is currently 2 steps East of Alice. The silver trophy is 3 steps South from her. The bronze trophy is 5 steps South from her. You see Alice walking 4 steps South.*

Both humans and our model strongly predict that Alice is headed towards the *bronze* trophy on this domain. In contrast, both the LLM baselines predict most strongly that Alice is heading towards the *silver trophy* (perhaps because it is closer to Alice), though in fact, humans tend to agree that because she has walked *past* this trophy, as an efficient actor, this is likely not her goal. We provide additional examples of similar efficient-planning situations in the Appendix.
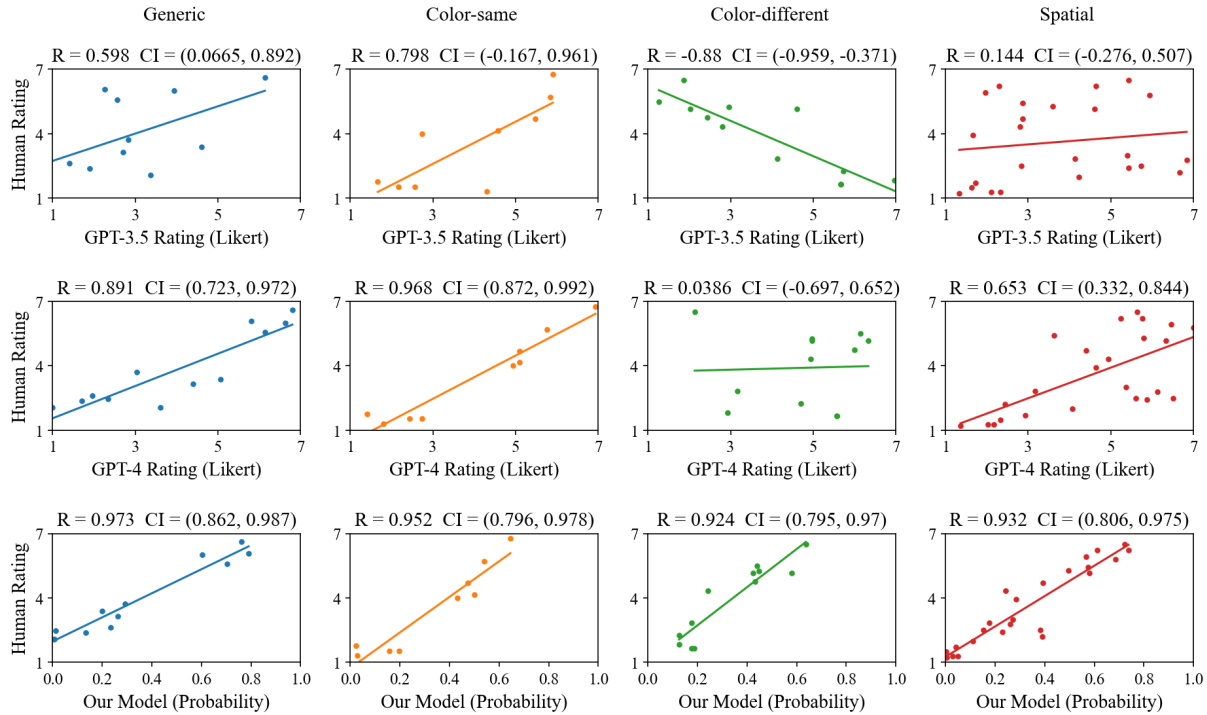
Figure 3. Correlations within specific task variants highlight more granular trends in model behavior. Our model (*Bottom*) correlates strongly with human judgements across all four task variants. In comparison, the LLM baselines generally show poorer performance on the more complex variants (the *Color-Different* variant, which introduces a new, counterintuitive rule about the world dynamics; and the *Spatial* variant, which requires reasoning specifically about agents as efficient path planners).

6. Conclusion and Future Directions

We conclude with several important avenues for future work. First, while we evaluate all LLMs using a simple prompting regime, future work should evaluate whether these results are consistent or improve with other prompting and evaluation methods such as chain-of-thought (Wei et al., 2022; Kojima et al., 2022), particularly on complex tasks such as those that invoke efficient spatial path planning, which prior work also finds to be an outstanding challenge for LLMs (Cohn & Hernandez-Orallo, 2023). It is also possible for LLMs to be biased by surface-level cues (e.g. in our stimuli, they might be biased towards gold over silver trophies, since the former are often considered more valuable). Future experiments should control for these potential biases by testing a wider diversity of prompts.

Second, while NIPE adopts a Bayesian approach to social inference, the process of language translation itself is performed in a non-Bayesian "bottom-up" manner using LLMs. Future work might explore a more fully Bayesian approach that models how language itself is generated from symbolic task representations, and perform posterior inference directly from language. This might involve using LLMs as likelihood models over language, as in Ying et al. (2023), and also incorporate LLMs as common sense priors over

representations of the world (Lew et al., 2020; Li et al., 2023). Among other benefits, this could enable more controlled and robust translation of language to code (Poesia et al., 2022; Lew et al., 2023).

Finally, we designed the gameshow domain to permit future extensions that describe a much broader range of social reasoning tasks that we could express linguistically. This includes scenarios that specifically invoke agent belief (Baker et al., 2017; Ullman, 2023); involve boundedly rational agents (Alanqary et al., 2021); and even involve multiple agents collaborating or competing (Kleiman-Weiner et al., 2016). In addition, future work could extend this domain to multimodal stimuli that include both linguistic information *and* demonstrations or images, e.g., to convey complex agent trajectories or environment layouts (Ying et al., 2023). The difficulty of the domain could also be increased, expanding the number of objects, actions, and planning steps in order to push the limits of language-to-code translation. By building out this architecture, we hope to fully reap the benefits of the both the neural and symbolic traditions of AI, enabling flexible social inferences from a wide range of naturally occurring linguistic stimuli, while producing outputs that are structured, coherent, and robust (Collins et al., 2022).

References

Aeronautiques, C., Howe, A., Knoblock, C., McDermott, I. D., Ram, A., Veloso, M., Weld, D., SRI, D. W., Barrett, A., Christianson, D., et al. PDDL — The Planning Domain Definition language. *Technical Report, Tech. Rep.*, 1998.

Alanqary, A., Lin, G. Z., Le, J., Zhi-Xuan, T., Mansinghka, V. K., and Tenenbaum, J. B. Modeling the mistakes of boundedly rational agents within a bayesian theory of mind. *arXiv preprint arXiv:2106.13249*, 2021.

Baillargeon, R., Scott, R. M., and Bian, L. Psychological reasoning in infancy. *Annual review of psychology*, 67: 159–186, 2016.

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

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

Bansal, G., Nushi, B., Kamar, E., Horvitz, E., and Weld, D. S. Is the most accurate ai the best teammate? optimizing ai for teamwork. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11405–11414, 2021.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., and Zhang, Y. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023.

Carroll, M., Shah, R., Ho, M. K., Griffiths, T., Seshia, S., Abbeel, P., and Dragan, A. On the utility of learning about humans for human-ai coordination. *Advances in neural information processing systems*, 32, 2019.

Cohn, A. G. and Hernandez-Orallo, J. Dialectical language model evaluation: An initial appraisal of the commonsense spatial reasoning abilities of llms. *arXiv e-prints*, pp. arXiv–2304, 2023.

Collins, K. M., Wong, C., Feng, J., Wei, M., and Tenenbaum, J. Structured, flexible, and robust: benchmarking and improving large language models towards more human-like behavior in out-of-distribution reasoning tasks. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44, 2022.

Csibra, G., Gergely, G., Bíró, S., Koos, O., and Brockbank, M. Goal attribution without agency cues: the perception of 'pure reason' in infancy. *Cognition*, 72(3):237–267, 1999.

Dennett, D. C. *The Intentional Stance*. MIT Press, 1981.

Evans, O., Stuhlmüller, A., and Goodman, N. Learning the preferences of ignorant, inconsistent agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

Gregory, P., Long, D., Fox, M., and Beck, J. C. Planning modulo theories: Extending the planning paradigm. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 22, pp. 65–73, 2012.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., and Tenenbaum, J. B. The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in cognitive sciences*, 20(8):589–604, 2016.

Kleiman-Weiner, M., Ho, M. K., Austerweil, J. L., Littman, M. L., and Tenenbaum, J. B. Coordinate to cooperate or compete: abstract goals and joint intentions in social interaction. In *CogSci*, 2016.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, 2022.

Lew, A. K., Tessler, M. H., Mansinghka, V. K., and Tenenbaum, J. B. Leveraging unstructured statistical knowledge in a probabilistic language of thought. In *Proceedings of the Annual Conference of the Cognitive Science Society*, 2020.

Lew, A. K., Zhi-Xuan, T., Grand, G., and Mansinghka, V. K. Sequential monte carlo steering of large language models using probabilistic programs. *arXiv preprint arXiv:2306.03081*, 2023.

Li, B. Z., Chen, W., Sharma, P., and Andreas, J. Lampp: Language models as probabilistic priors for perception and action. *arXiv e-prints*, pp. arXiv–2302, 2023.

Lian, B., Xue, W., Lewis, F. L., and Chai, T. Online inverse reinforcement learning for nonlinear systems with adversarial attacks. *International Journal of Robust and Nonlinear Control*, 31(14):6646–6667, 2021.

Liu, B., Jiang, Y., Zhang, X., Liu, Q., Zhang, S., Biswas, J., and Stone, P. Llm+ p: Empowering large language models with optimal planning proficiency. *arXiv preprint arXiv:2304.11477*, 2023a.

Liu, B., Jiang, Y., Zhang, X., Liu, Q., Zhang, S., Biswas, J., and Stone, P. Llm+p: Empowering large language models with optimal planning proficiency, 2023b.

Lyu, Q., Havaldar, S., Stein, A., Zhang, L., Rao, D., Wong, E., Apidianaki, M., and Callison-Burch, C. Faithful chain-of-thought reasoning. *arXiv preprint arXiv:2301.13379*, 2023.

Mozannar, H. and Sontag, D. Consistent estimators for learning to defer to an expert. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7076–7087. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/mozannar20b.html.

OpenAI. Gpt-4 technical report, 2023.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Poesia, G., Polozov, O., Le, V., Tiwari, A., Soares, G., Meek, C., and Gulwani, S. Synchromesh: Reliable code generation from pre-trained language models. *arXiv preprint arXiv:2201.11227*, 2022.

Ramachandran, D. and Amir, E. Bayesian inverse reinforcement learning. In *IJCAI*, volume 7, pp. 2586–2591, 2007.

Saxe, R. Uniquely human social cognition. *Current opinion in neurobiology*, 16(2):235–239, 2006.

Shah, R., Gundotra, N., Abbeel, P., and Dragan, A. On the feasibility of learning, rather than assuming, human biases for reward inference. In *International Conference on Machine Learning*, pp. 5670–5679. PMLR, 2019.

Shapira, N., Levy, M., Alavi, S. H., Zhou, X., Choi, Y., Goldberg, Y., Sap, M., and Shwartz, V. Clever hans or neural theory of mind? stress testing social reasoning in large language models. *arXiv preprint arXiv:2305.14763*, 2023.

Shneiderman, B. *Human-centered AI*. Oxford University Press, 2022.

Shu, T., Bhandwaldar, A., Gan, C., Smith, K., Liu, S., Gutfreund, D., Spelke, E., Tenenbaum, J., and Ullman, T. Agent: A benchmark for core psychological reasoning. In *International Conference on Machine Learning*, pp. 9614–9625. PMLR, 2021.

Silver, T., Dan, S., Srinivas, K., Tenenbaum, J. B., Kaelbling, L. P., and Katz, M. Generalized planning in pddl domains with pretrained large language models, 2023.

Tomasello, M. *Origins of human communication*. MIT press, 2010.

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

Valmeekam, K., Olmo, A., Sreedharan, S., and Kambhampati, S. Large language models still can't plan (a benchmark for llms on planning and reasoning about change). *arXiv preprint arXiv:2206.10498*, 2022.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E. H., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022.

Wellman, H. M. *The child's theory of mind*. The MIT Press, 1992.

Wilder, B., Horvitz, E., and Kamar, E. Learning to complement humans. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pp. 1526–1533, 2021.

Wong, L., Grand, G., Lew, A. K., Goodman, N. D., Mansinghka, V. K., Andreas, J., and Tenenbaum, J. B. From word models to world models: Translating from natural language to the probabilistic language of thought, 2023.

Ying, L., Zhi-Xuan, T., Mansinghka, V., and Tenenbaum, J. Inferring the goals of communicating agents from actions and instructions. *ICML Workshop on Theory of Mind in Communicating Agents (To Appear)*, 2023.

Zhang, C. E., Wong, L., Grand, G., and Tenenbaum, J. B. Grounded physical language understanding with probabilistic programs and simulated worlds. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 45, 2023.

Zhi-Xuan, T. *PDDL. jl: An Extensible Interpreter and Compiler Interface for Fast and Flexible AI Planning*. PhD thesis, Massachusetts Institute of Technology, 2022.

Zhi-Xuan, T., Mann, J., Silver, T., Tenenbaum, J., and Mansinghka, V. Online bayesian goal inference for boundedly rational planning agents. *Advances in neural information processing systems*, 33:19238–19250, 2020.

Ziebart, B. D., Maas, A. L., Bagnell, J. A., Dey, A. K., et al. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pp. 1433–1438. Chicago, IL, USA, 2008.

A. Appendix

A.1. Acknowledgments

A.2. Human linguistic inverse planning experiment

We include additional details on our human experiment. For all scenarios, participants were asked "Now, considering each of the possible trophies on the course, please respond on the 1-7 scale: how likely is it that the participant is trying to reach this trophy?" Likert scale options were as follows:

- "1 (Definitely not this trophy)"

- "2 (Very unlikely)"

- "3 (Somewhat unlikely)"

- "4 (Equally likely or unlikely)"

- "5 (Somewhat likely)"

- "6 (Very likely)"

- "7 (Definitely this trophy)"

Note, due to a mismatch, some participants in the `Generic` condition saw one stimuli from `Different-Colored Keys` and vice versa. We reversed fixed this partway through the study.

A.3. LLM Baselines Experiment

We use the gpt-4-0314 model (i.e., a fixed, non-updatable checkpoint from March 14 for maximal reproducibility) in the Chat Completion API. For each stimuli, we sample until we get 20 well-formed responses, rejecting samples that are not in the correct format.

Example LLM Prompt for World Configuration

```
Input:
There are three trophies placed throughout this obstacle course: gold, silver, or bronze.
The gold trophy is behind a green door.
The silver trophy is behind a yellow and red door.
The bronze trophy is behind a red door.

Keys to unlock the doors are placed around the obstacle course. On this course, each door
can be unlocked by any key as long as it is of a different color than the door.

You see Alice picking up a red key. What would be the likely trophy Alice wants to get?"


output:
{"agent" : ["Alice"],
"goals" : ["gold", "silver", "bronze"],
"locations" : {"gold" : "Room A", "silver" : "Room B", "bronze" : "Room C"},
"obstacles" : {"Room A" : ["G"], "Room B" : ["Y", "R"],"Room C" : ["R"] },
"keys" : ["g","y","r"],
```

```
"max_obstacle" : 2,
"keys_per_door": 1,
"len_key" : 3,
"goal_count" : 3,
"observation_type": "has_objects",
"observation": "(exists (?k - key) (and (has Alice ?k) (iscolor ?k red)))"}
```

Example LLM Prompt for PDDL Domain Definition

In this example we constrain the model such that the LLM can only modify the unlock action of the PDDL domain definition. Our goal is to show that with few-shot prompting in the color-same condition, the LLM can generalize to generate action definition for color-different condition.

```
input:

Alice is playing a treasure game to pick up exactly one trophy.
There are 3 trophies: gold, silver and bronze, located in room A, room C and room B.
There is 1 yellow and 1 green lock for Room A, a red lock for room B and a
yellow lock for room C.

Alice needs to unlock the locks to enter the rooms. Each lock can be unlocked
with a key with the same color. Each key can only be used once.

You see Alice go over to pick up a red key.

It's unclear what Alice would do next.

output:

(:action unlock
 :parameters (?a - agent ?k - key ?d - door)
 :precondition (and (has ?a ?k) (locked ?d)
                    (exists (?c - color) (and (iscolor ?k ?c) (iscolor ?d ?c)))
                    (or (and (= (xloc ?a) (xloc ?d)) (= (- (yloc ?a) 1) (yloc ?d)))
                        (and (= (xloc ?a) (xloc ?d)) (= (+ (yloc ?a) 1) (yloc ?d)))
                        (and (= (- (xloc ?a) 1) (xloc ?d)) (= (yloc ?a) (yloc ?d)))
                        (and (= (+ (xloc ?a) 1) (xloc ?d)) (= (yloc ?a) (yloc ?d)))))
 :effect (and (not (has ?a ?k)) (not (locked ?d)) )
)
```

A.4. An example of LLM baseline result

Here we provide an example of GPT-4's response in the baseline condition on the *Color-different* task variant.

```
input:

There are three trophies placed throughout this obstacle course:
gold, silver, or bronze.

The gold and silver trophies are both located behind a yellow door.
To go to the bronze trophy, you need to go through a red door.

Keys to unlock the doors are placed around the obstacle course.
```

```
On this course, each door can be unlocked by any key that is a different
color than the door. For example, yellow keys cannot unlock yellow doors,
but can unlock doors in green or red.

In this game, players lose points if they pick up more keys than needed.
You may assume that players will use all the keys they pick up.

Alice is a participant on this course. You see Alice pick up a yellow key.

It's unclear what Alice would do next.

How likely is Alice picking up the trophies? Please rate from 1 to 7
and explain.

output:

Gold Trophy: 6
Silver Trophy: 6
Bronze Trophy: 1

Explanation:

Since Alice has picked up a yellow key, she can now unlock either
the gold or silver trophy, since both are located behind a yellow
door. This makes it quite likely that she will pick up one of those
two trophies (rating them both at 6).

However, picking up a yellow key means she cannot unlock the red
door, which leads to the bronze trophy. Therefore, it is very unlikely
that Alice will pick up the bronze trophy (rating it 1).
```

It is clear from this example that GPT-4 shows unsatisfactory understanding of the setup of the *Color-different* variant. It is stated that yellow keys cannot unlock yellow doors, so it is unlikely that Alice would pick up the gold or silver trophy, but GPT-4 still judges these two possibilities to be highly likely. On the other hand, both humans and our NIPE model assign significantly more likelihood to the bronze trophy (and roughly equal low likelihoods to the gold and silver trophies) in this example.

A.5. Sample stimuli for each experimental condition

Generic:

There are three trophies placed throughout this obstacle course: gold, silver, or bronze.

To get to the gold trophy, you need to unlock 2 different doors. To get to the silver trophy, you need to unlock 1 door. The bronze trophy isn't behind a door.

Keys to unlock the doors are placed around the obstacle course. On this course, each door has exactly 1 lock on it. All of the keys on this course work on all of the locks. However, each key can only be used one time.

Alice is a participant on this course. You see Alice go over and pick up two keys.

Color-same:

There are three trophies placed throughout this obstacle course: gold, silver, or bronze.

The gold and silver trophy trophies are both located behind a yellow door. To get to the yellow door, you first need to go through a green door. To go to the bronze trophy, you need to go through a red door.

Keys to unlock the doors are placed around the obstacle course. On this course, each door is unlocked by a key of the corresponding color.

In this game, players lose points if they pick up more keys than needed. You may assume that players will use all the keys they pick up.

Alice is a participant on this course. You see Alice pick up a yellow key.

Color-different:

There are three trophies placed throughout this obstacle course: gold, silver, or bronze, which are placed in rooms labeled A, B, and C, respectively.

Room A is behind a door with a yellow and green lock on it. Room B is behind a door with a red lock on it. Room C is behind a door with its own yellow lock on it.

Keys to unlock the doors are placed around the obstacle course. On this course, each door can be unlocked by any key that is a different color than the door. For example, yellow keys cannot unlock yellow doors, but can unlock green and red doors

In this game, players lose points if they pick up more keys than needed. You may assume that players will use all the keys they pick up.

Alice is a participant on this course. You see Alice pick up a yellow key.

It's unclear what Alice would do next.

Spatial:

There are three trophies placed throughout this obstacle course: gold, silver, or bronze. Alice is a participant on this course.

The gold trophy is currently 1 step East of Alice. The silver trophy is 3 steps South from her. The bronze trophy is 3 steps East from the silver trophy.

You see Alice walking 3 steps East.