Modeling Open-World Cognition as On-Demand Synthesis of Probabilistic Models

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

When faced with novel situations, people can marshal relevant considerations from a wide range of background knowledge and use these for inference and prediction. How do we draw in globally relevant information and reason over it coherently? We explore the hypothesis that people reason by constructing structured but small, ad-hoc mental models on the fly, tailored to novel situations. We propose a computational implementation of this idea – a "Model Synthesis Architecture" (MSA) – using language models to parameterize global, relevance-based retrieval of variables, and probabilistic programs to implement bespoke, coherent world models. We evaluate our MSA, along with ablations and baselines, as a model of human judgments across a sequence of experiments that requires progressively more open-ended and open-world reasoning about situations described in natural language. Across all experiments, the MSA captures human judgments, and outperforms the base LM alone – suggesting that MSAs offer a path towards capturing coherent human reasoning in open-ended domains.

1 Introduction

An influential idea in cognitive science holds that people reason and plan using mental models, or structured internal representations that mirror aspects of the world [12, 31, 19]. In this view, people draw on structured mental models to maintain consistent beliefs about current world states, integrate new information into their beliefs, and evaluate the plausibility of alternative hypotheses or possible futures. This idea also appears throughout classic and recent work in AI, in theoretical proposals [28, 39, 2] and empirical investigations (e.g., [42]) predicated on the idea that intelligent systems should reason and plan using structured internal representations of the causal systems and environments they operate over.

In cognitive science, Bayesian modeling has found significant empirical support for a version of the "mental modeling" hypothesis, showing that human judgments across a wide variety of tasks are well-modeled by inference and decisions in causal, probabilistic models (e.g. about physical predictions [4, 26], causal learning [23], and social reasoning [3, 30], to name just a few). However, while these models are predictive of human judgments and learning in each of these settings, they remain importantly limited in that each model operates *only* in the limited scope for which it was designed. Any given model can provide inferences over the variables it represents, but cannot handle novel considerations that were not part of the initial model specification. People, by contrast, are 'open-world' reasoners. We regularly reason about novel questions that draw on a highly varied set of things we know about the world, any of which are likely to be missing from any given mental model. To date, Bayesian modeling has left it unclear how such modeling approaches could scale to explain the simultaneous flexibility and coherence of human reasoning in general. This scalability

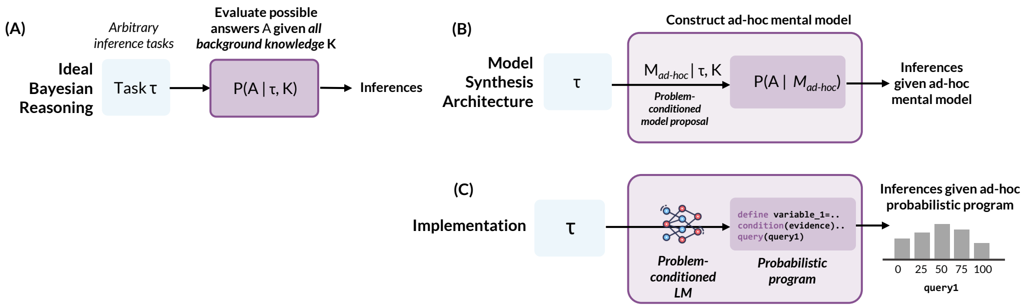

(A)
Arbitrary inference tasks
Evaluate possible answers A given *all background knowledge* K
Ideal Bayesian Reasoning
Task τ
P(A | τ, K)
Inferences
(B)
Construct ad-hoc mental model
Model Synthesis Architecture
τ
M_ad-hoc | τ, K
Problem-conditioned model proposal
P(A | M_ad-hoc)
Inferences given ad-hoc mental model
(C)
Implementation
τ
define variable_1=..
condition(evidence)..
query(query1)
Problem-conditioned LM
Probabilistic program
Inferences given ad-hoc probabilistic program
0 25 50 75 100
query1

Figure 1: (A) Idealized Bayesian reasoning about arbitrary tasks τ given global background knowledge K raises problems of computational tractability. (B) *Model Synthesis Architectures* use global relevance functions to construct ad-hoc, structured mental models for locally coherent reasoning. (C) We implement an MSA using LMs to parameterize global relevance functions and Probabilistic Programs to construct ad-hoc probabilistic models.

challenge is also one of the most significant barriers to seeing structured Bayesian modeling as a general approach to building AI systems that capture these hallmarks of human reasoning.

So, how do people reason in *locally coherent* ways in any given context, while drawing *globally* on potentially relevant considerations across their background knowledge and beliefs? In this paper, we explore the hypothesis that human minds implement "Model Synthesis Architectures" (MSAs, Figure 1), or architectures that construct small, ad-hoc mental models on the fly in response to task demands [7]. By reasoning within small models, MSAs can deliver local coherence over the variables they explicitly represent, while the ability to synthesize arbitrary models as needed allows the architecture to reason and plan in open-ended environments, where the relevant considerations are not fixed in advance. We implement a concrete instance of an MSA (Figure 1C) using Probabilistic Programming Languages (PPLs) [21, 5, 8, 14] to express individual models as probabilistic programs, and using a neurally-guided program synthesis procedure, constituted by structured calls to a Language Model (LM), to construct relevant mental models. Our goal in combining these is to build a system that, like human cognition, can operate in the open-world setting while still delivering the natively coherent reasoning of structured probabilistic models and addressing concerns about the fragility of internal, "world model"-like representations in language models alone (e.g., [42, 36]).

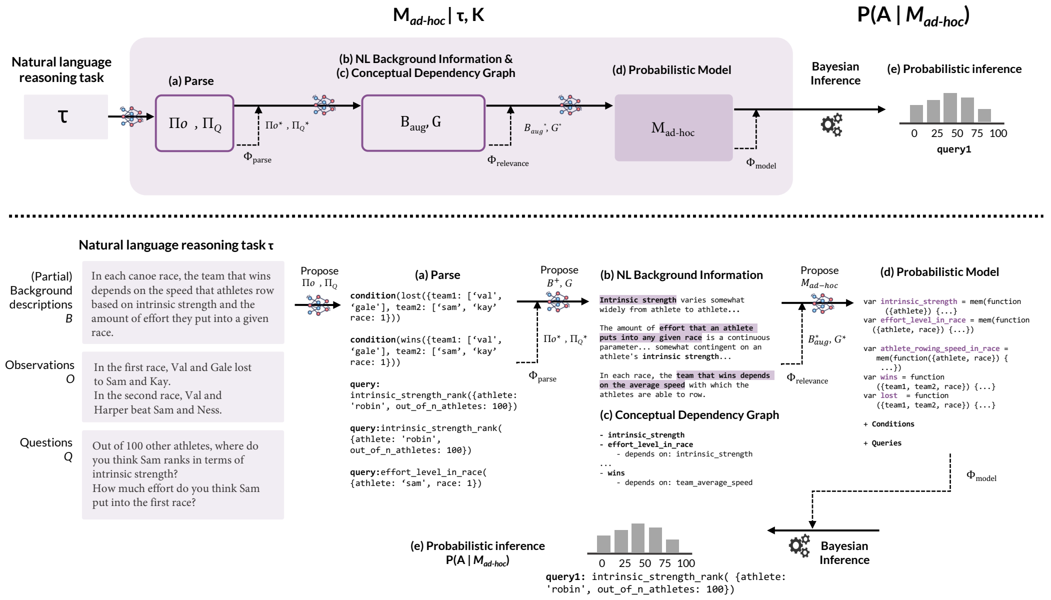
We study this MSA implementation empirically and compare it with human ad-hoc reasoning, as well as pure LM and PPL baselines using a domain of natural language inference tasks designed to test generalization and coherence. We design a sequence of experiments to test coherent open-world reasoning. We first evaluate how people and models reason on a relatively controlled set of natural language inference problems, then construct successive experiments that demand progressively more generalization to novel variables while drawing on more distant background knowledge. Across all experiments, we find that human reasoning is well-captured by our Model Synthesis Architecture, which provides a better match to human judgments than LM-only baselines and model ablations. This represents a proof of concept that neural language modeling and structured probabilistic modeling can be interleaved to explain people's ability to reason in ways that are globally relevant and locally coherent in an open-world setting.

2 Model Synthesis Architectures

We consider the general problem of inferring answers A to an arbitrary inference or prediction problem τ. In the idealized Bayesian inference setting (Figure 1, top), drawing these inferences involves conditioning on the information in the specific problem τ in light of *all of the reasoner's prior background knowledge* K to produce answers:

$$P(A \mid \tau, K). \tag{1}$$

Computing this will often be prohibitively costly, as probabilistic inference is intractable in the general case, and wasteful, since on any one occasion, it's likely that only a small portion of what the reasoner actually knows will matter for the question at hand. Instead of computing this full conditional, we propose that a reasoner "marshals" only a subset of their background knowledge (K')



In each canoe race, the team that wins depends on the speed that athletes row based on intrinsic strength and the amount of effort they put into a given race.

In the first race, Val and Gale lost to Sam and Kay.
In the second race, Val and Harper beat Sam and Ness.

Out of 100 other athletes, where do you think Sam ranks in terms of intrinsic strength?
How much effort do you think Sam put into the first race?

```
condition(lost({team1: [‘val',
‘gale'], team2: [‘sam’, ‘kay’
race: 1}))

condition(wins({team1: [‘val',
‘gale'], team2: [‘sam’, ‘kay’
race: 1}))

query:
intrinsic_strength_rank({athlete:
'robin', out_of_n_athletes: 100})

query:intrinsic_strength_rank(
{athlete: 'robin',
out_of_n_athletes: 100})

query:effort_level_in_race(
{athlete: ‘sam', race: 1})
```

```
Intrinsic strength varies somewhat
widely from athlete to athlete...

The amount of effort that an athlete
puts into any given race is a continuous
parameter... somewhat contingent on an
athlete's intrinsic strength...

In each race, the team that wins depends
on the average speed with which the
athletes are able to row.
```

```
- intrinsic_strength
- effort_level_in_race
    - depends on: intrinsic_strength
...
- wins
    - depends on: team_average_speed
```

```
var intrinsic_strength = mem(function
    ({athlete}) {...}
var effort_level_in_race = mem(function
  ({athlete, race}) {...})

var athlete_rowing_speed_in_race =
   mem(function({athlete, race}) {
    ...})
var wins = function
   ({team1, team2, race}) {...}
var lost  = function
   ({team1, team2, race}) {...}

+ Conditions

+ Queries
```

```
query1: intrinsic_strength_rank( {athlete:
'robin', out_of_n_athletes: 100})
```

Figure 2: **(Top)** Schematic overview of the MSA implementation, which sequentially constructs $M_{\text{ad-hoc}}$ from input natural language tasks τ through interleaved LM-guided generation steps, and scoring steps using intermediate scoring functions Φ. **(Bottom)** Detailed overview of the MSA implementation. Given a task τ as a (potentially partial) background description B, observations O, and questions Q, we sequentially construct $M_{\text{ad-hoc}}$ through parsing the current inputs, retrieval of additional background knowledge in language, proposal of a conceptual dependency graph, and finally synthesis of a formal probabilistic model in which we compute *model-based Bayesian inferences*.

that is relevant to the problem at hand, such that:

$$P(A|\tau, K') \approx P(A \mid \tau, K). \tag{2}$$

In particular, to draw coherent inferences, we propose that reasoners use this reduced set of background knowledge to construct a *context-specific mental model* ($M_{\text{ad-hoc}}$) which they use to perform probabilistic inferences (Figure 1, middle), assuming that:

$$P(A|M_{\text{ad-hoc}}) \approx P(A|\tau, K'). \tag{3}$$

We call any system that implements this overarching cognitive hypothesis a *Model Synthesis Architecture*, as it decomposes reasoning about an arbitrary problem into two distinct computational subproblems: (1) a problem-conditioned, *ad-hoc model synthesis* step to construct $M_{\text{ad-hoc}}$, and then (2) a step computing *Bayesian model-based inferences* to answer the question conditional on the constructed model, i.e. computing $P(A|M_{\text{ad-hoc}})$.

The formal nature of each of these subproblems differs. As in *resource-rational* framings [34], we propose that reasoners treat ad-hoc model synthesis as an optimization problem, selecting representations that they believe will be *useful* for reasoning about a problem:

$$\text{argmax}_{i \in k_{model}} \Phi(M^i_{\text{ad-hoc}}, \tau) \tag{4}$$

based on various model evaluation functions Φ (e.g., trading off between computational costs of inference in the model with expected accuracy for a set of queries) over a set of k_{model} sampled models. In contrast to model synthesis, reasoning and planning with a synthesized model might look like optimization, inference, or deduction, depending on the problem and synthesized model. We focus on probabilistic inference, where models represent structured priors or *conceptions* of the relevant variables and dependencies for the problem at hand, cf. [20, 22].

In this paper, we consider a subset of τ, the space of natural language probabilistic reasoning problems defined by a tuple (B, O, Q), where B is a (potentially partial and underspecified) set of *background* variable descriptions $b_1, ..., b_N$ about a situation at hand (e.g., someone trying to predict upcoming sports tournaments in a bracket might mention factors like injuries or training that they believe should be considered); O is a set of *observations* $o_1, ..., o_N$ providing evidence that bears on those variables (e.g., observations about which teams have previously won or lost in the tournament); and Q is a set of *questions* $q_1, ..., q_N$ that single out particular queries to answer given the evidence (e.g., specific prediction questions about which teams will win in an upcoming match).

2.1 Representing and synthesizing ad-hoc models

In this section, we describe a concrete MSA implementation in which ad-hoc models are represented as task-specific *probabilistic programs*. Each probabilistic program represents models as a tuple $M_{\text{ad-hoc}} = (\Pi_B, \Pi_O, \Pi_Q)$, in which Π_B are set of stochastic function definitions that formalize a causal prior over relevant background variables, by defining their distributional form and causal dependencies; Π_O are a set of observed constraints over defined variables which condition belief under the prior; and Π_Q are query expressions over the defined variables which define targets for Bayesian inference under the conditioned probabilistic model.

Our concrete implementation then frames model synthesis as *LM-guided probabilistic program synthesis*, using LMs to parameterize a search procedure over programs conditioned on an input task, and to parameterize a set of model evaluation functions Φ. This implementation ultimately answers queries in synthesized models as $P(A|M_{\text{ad-hoc}})$, using automatic Bayesian inference procedures defined generally over the probabilistic programming language. We present an implementation that approximates this optimization over models via a *sequentially staged* synthesis process, interleaving structured steps of partial model generation and evaluation. Interleaving generation and evaluation allows us to focus future generation stages on outputs that evaluate highly under components of Φ so far, providing efficiency gains. We briefly overview these stages here; additional details on each stage can be found in the supplement:

Parse (Figure 2a): We first parse the current natural language inputs ($\tau = B, O, Q$) into a set of candidate probabilistic program condition and query expressions (Π_O, Π_Q) to be passed on to future model synthesis stages. (We do not yet parse the background B, as we expand on that in the future stages.) Specifically, we use an LM that has been prompted to parse each sentence of input natural language observations into a corresponding formal expression (π_O) intended to condition a probabilistic model with observed constraints on latent variables, and questions into query expressions (π_Q) that define target variables for inference in that model. We sample proposed parses from an LM conditioned on the input task (see Appendix for prompting details):

$$\Pi_O, \Pi_Q \sim P_{\text{LM}}(\Pi_O, \Pi_Q|\tau) \tag{5}$$

We then score each sampled parse according to an evaluation function Φ_{parse} (also defined using an LM prompted with example parses). We generate k_{parse} candidates and greedily select the best conditioned on the input task:

$$\Pi_O^*, \Pi_Q^* = \text{argmax}_{i \in k_{parse}} \Phi_{parse}(\Pi_O^i, \Pi_Q^i|\tau) \tag{6}$$

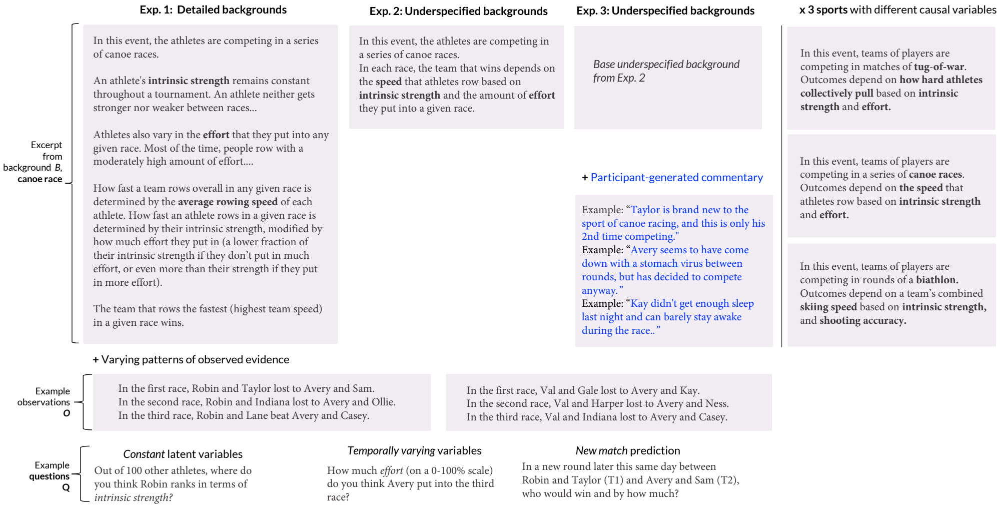
This best parse (we use the * notation throughout to refer to the best scoring generations, from a set of candidates, with respect to a utility function Φ) is then passed on to the next stage.

Relevant Natural Language Background Description (Figure 2b): Next, we retrieve candidates for additional, relevant background knowledge details ($B^+ = \{b_1^+, b_2^+, ...b_{N'}^+\}$). These will be combined with the initial (potentially underspecified) input background B to yield an augmented natural language description $B_{aug} = B \bigcup B^+$ that is intended to fully specify, in explicit detail, latent relevant variables for reasoning about the task at hand. We sample these additional background details from an LM prompted to name relevant variables and explicitly describe their causal relationship on other variables, conditioned on the previous stages of generation:

$$B_{aug} \sim P_{\text{LM}}(b_1^+, b_2^+, ...b_{N'}^+|\Pi_O^*, \Pi_Q^*, \tau). \tag{7}$$

Conceptual Dependency Graph (Figure 2c): Jointly with generating B_{aug}, we also generate a conceptual dependency graph G which explicitly summarizes the dependencies between all variables in B_{aug}. We jointly score B_{aug} and G using an LM-parameterized evaluation function $\Phi_{relevance}$ (defined using an LM prompted with example retrieved variables and graphs). We generate $k_{relevance}$ candidates and greedily select the best:

$$B_{aug}^*, G^* = \text{argmax}_{i \in k_{relevance}} \Phi_{relevance}(\{B_{aug}, G\}^i|\Pi_O^*, \Pi_Q^*, \tau) \tag{8}$$


Exp. 1: Detailed backgrounds
Exp. 2: Underspecified backgrounds
Exp. 3: Underspecified backgrounds
x 3 sports with different causal variables
Excerpt from background *B*, canoe race

In this event, the athletes are competing in a series of canoe races.

An athlete's **intrinsic strength** remains constant throughout a tournament. An athlete neither gets stronger nor weaker between races...

Athletes also vary in the **effort** that they put into any given race. Most of the time, people row with a moderately high amount of effort....

How fast a team rows overall in any given race is determined by the **average rowing speed** of each athlete. How fast an athlete rows in a given race is determined by their intrinsic strength, modified by how much effort they put in (a lower fraction of their intrinsic strength if they don't put in much effort, or even more than their strength if they put in more effort).

The team that rows the fastest (highest team speed) in a given race wins.

In this event, the athletes are competing in a series of canoe races.
In each race, the team that wins depends on the **speed** that athletes row based on **intrinsic strength** and the amount of **effort** they put into a given race.

Base underspecified background from Exp. 2

Example: "Taylor is brand new to the sport of canoe racing, and this is only his 2nd time competing."
Example: "Avery seems to have come down with a stomach virus between rounds, but has decided to compete anyway."
Example: "Kay didn't get enough sleep last night and can barely stay awake during the race.."

In this event, teams of players are competing in matches of **tug-of-war**. Outcomes depend on **how hard athletes collectively pull** based on **intrinsic strength** and **effort.**

In this event, teams of players are competing in a series of **canoe races**. Outcomes depend on **the speed** that athletes row based on **intrinsic strength** and **effort.**

In this event, teams of players are competing in rounds of a **biathlon.** Outcomes depend on a team's combined **skiing speed** based on **intrinsic strength,** and **shooting accuracy.**

In the first race, Robin and Taylor lost to Avery and Sam.
In the second race, Robin and Indiana lost to Avery and Ollie.
In the third race, Robin and Lane beat Avery and Casey.

In the first race, Val and Gale lost to Avery and Kay.
In the second race, Val and Harper lost to Avery and Ness.
In the third race, Val and Indiana lost to Avery and Casey.

Out of 100 other athletes, where do you think Robin ranks in terms of *intrinsic strength?*

How much *effort* (on a 0-100% scale) do you think Avery put into the third race?

In a new round later this same day between Robin and Taylor (T1) and Avery and Sam (T2), who would win and by how much?

Figure 3: Experiment overview for the three natural language reasoning experiments.

Probabilistic Model (Figure 2d): We now generate a full symbolic probabilistic model $M_{\text{ad-hoc}}$. We sample candidate models as probabilistic programs from an LM prompted with examples of the preceding steps of generation and corresponding programs:

$$M_{\text{ad-hoc}} \sim P_{\text{LM}}(M_{\text{ad-hoc}} | B^*_{aug}, G^*, \Pi^*_O, \Pi^*_Q, \tau). \tag{9}$$

We evaluate the sampled model for formal validity Φ_{model} (we use a simple Boolean function that just returns executable models) and finally return the best model as:

$$M_{\text{ad-hoc}}* = \text{argmax}_{i \in k_{program}} \Phi_{\text{model}}(M^i_{\text{ad-hoc}}). \tag{10}$$

Probabilistic Inference (Figure 2e): Finally, we compute probabilistic inferences in $M_{\text{ad-hoc}}$ using general Bayesian inference algorithms defined over the probabilistic programming language. We return inferences in the queried and conditioned model.

$$P(A | M_{\text{ad-hoc}}*). \tag{11}$$

as a joint distribution over the set of answers A corresponding to input questions.

3 Natural Language Reasoning Experiments

To evaluate flexible ad-hoc reasoning in people and models, we construct a domain of *natural language inference problems*. We then design three experiments around this domain that require people and models to bring to bear progressively more background information to reason about the problem at hand. Here we briefly overview the domain and progression of experiments; additional information on all stimuli and experiments can be found in our supplement.

3.1 Domain – Model Olympics vignettes

Our domain is a set of vignettes about three different sporting events – tug-of-war, canoe-racing, and biathlon – each with distinct causal structures and variables. Each vignette includes a set of observations about teams of athletes competing in a specific set of matches (e.g. *In the first race, Robin and Taylor beat Avery and Sam*), and a palette of 8 different questions that require inferring latent variables and new predictions. Of these 8 questions, 3 are always about a *constant* latent variable (e.g the underlying *strength* of the athletes), 3 are about a *temporally varying* latent variable (eg. the amount of effort an athlete puts into a given match), and 2 require making predictions about future matches. We construct 7 underlying vignette templates designed to probe different patterns of evidence (e.g. *anomalous loss* cases where an otherwise strong team happens to lose). In total, we design 21 vignettes (7 templates x 3 sports).

3.2 Human and model experiments

Experiment 1 (detailed background context) tests ad-hoc reasoning about arbitrary collections of variables when they are described explicitly for a given situation. Vignettes are presented with detailed linguistic background descriptions (Figure 3, left) based on the gold, hand-crafted probabilistic models for each sport. These background descriptions spell out the functional form of relevant variable distributions. $N_{E1} = 78$ participants from Prolific judged a random sample of two vignettes from each of the sports.

Experiment 2 (under-specified background context) tests ad-hoc reasoning when some relevant variables are briefly described or implied in language, but most of the relevant intermediate details must be filled in from a participants' background knowledge. Vignettes are presented with brief and under-specified background descriptions (Figure 3, center) that name the key variables for each sport, but do not explicitly spell out the functional form of their underlying distributions or how these variables interact to produce observed outcomes. $N_{E2} = 80$ participants from Prolific judged the same batches of vignettes as in Experiment 1.

Experiment 3 (participant-generated novel details) tests how people flexibly incorporate arbitrary evidence into ad-hoc reasoning, by introducing uncontrolled new variables from outside the scope of our original domain. To do this, we extend the under-specified vignettes from Experiment 2 with new participant-generated details ("sports commentary", Figure 3, right), from naive participants prompted to come up with new, relevant observations that would have changed their own predictions about a particular future match (eg. to make a given outcome more or less likely).Participant-generated details were collected from $N_{E3,a} = 20$ naive human subjects in a separate elicitation task. This experiment used a smaller set of 9 total vignettes across two sports (tug-of-war and canoe racing). In the main judgment task with the extended vignettes, $N_{E3,b} = 20$ participants judged all 9 vignettes.

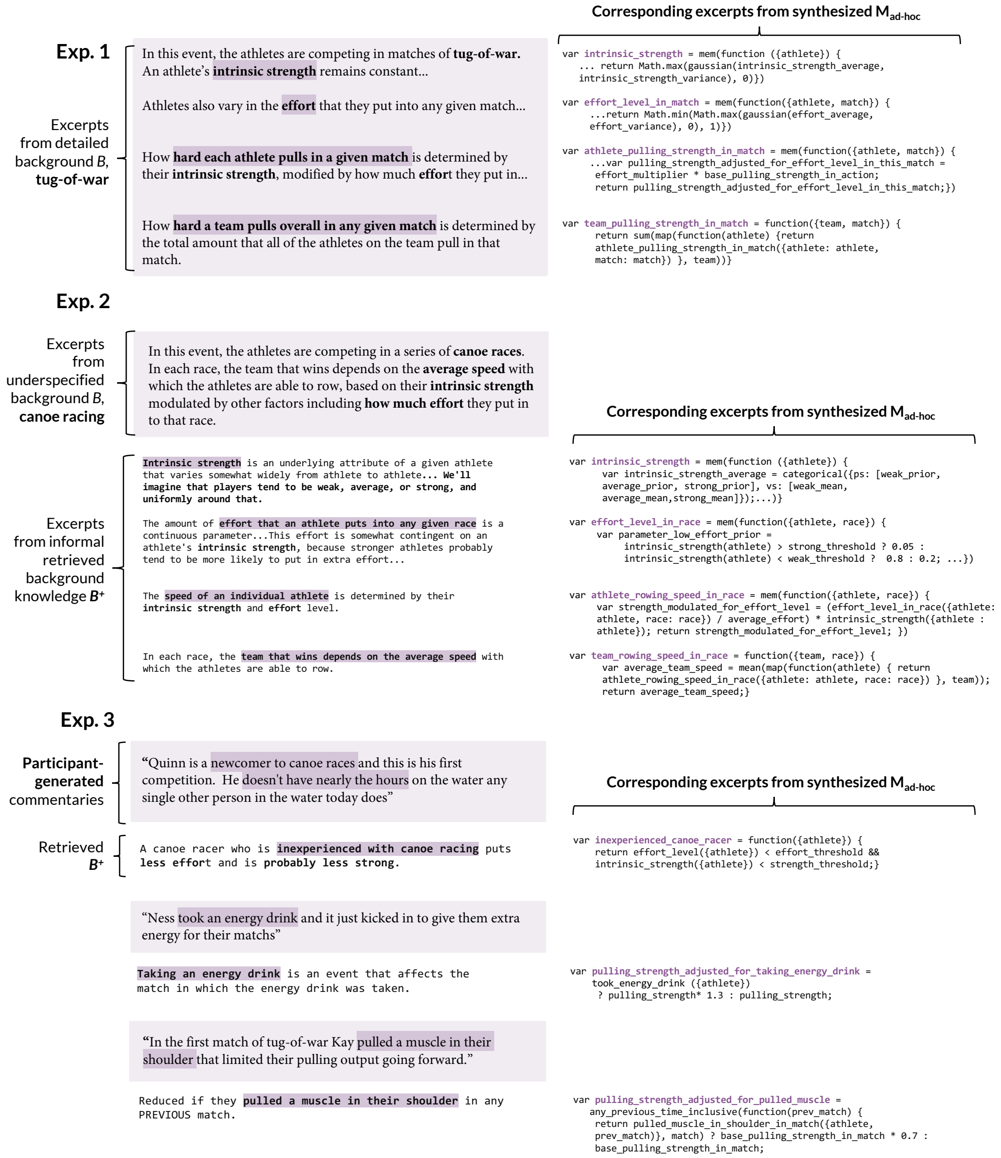
We instantiate our MSA architecture using `Llama-3.1-70B` as our base LM for all parsing, code synthesis, and LM-based evaluations; and WebPPL [1] as our probabilistic programming language. Across all three experiments, we elicit simulated judgments from the MSA and alternatives in the form of estimated posteriors for the palette of questions, conditioned on the vignettes. We compare the MSA judgments to those of several alternative models:

- **Gold symbolic models:** For Exp. 1 and Exp. 2, we estimate posteriors using the hand-designed, gold symbolic models constructed for each of the three sports
- **Large language models (direct and CoT)**: We also compare to two LM-only alternatives using the base LM model (Llama-3.1-70B): a"direct" response setting, where we prompt the LM to directly answer all questions for each vignette via feedforward generation, and a chain-of-thought (CoT) setting [45].

4 Results

We compare human judgments to all models using both **correlational measures** (R^2) between people and models, computed between the mean judgments across all participants (combining all participant clicks for each query) and the mean judgments for each model (combining all simulated participant samples for each query); and **distributional measures**, computed as the Wasserstein Distance (WD) metric, to compare the similarity in probability distributions over inferred variables – **lower** Wasserstein Distances between model judgments and human judgments mean that the distributions are *more similar* to one another.

Key finding 1: People's reasoning is generally consistent with Bayesian inference in ad-hoc probabilistic models. Figure 5 and Figure 6 (A-C) shows that across all three experiments, inferences in the probabilistic models synthesized by our MSA are generally well-correlated with human judgments. This result is also borne out by our distributional analyses. Figure 6 (D-F) shows that the MSA captures not just the average human judgment, but often the distribution of predictions. Under this metric, comparison to the human-human baseline again shows that the distribution of MSA inferences is about as similar to human judgments (*purple*) as humans are to each other (*blue*).


Corresponding excerpts from synthesized $M_{\text{ad-hoc}}$
Exp. 1
Excerpts from detailed background *B*, **tug-of-war**

In this event, the athletes are competing in matches of **tug-of-war.** An athlete's **intrinsic strength** remains constant...

Athletes also vary in the **effort** that they put into any given match...

How **hard each athlete pulls in a given match** is determined by their **intrinsic strength**, modified by how much **effor**t they put in...

How **hard a team pulls overall in any given match** is determined by the total amount that all of the athletes on the team pull in that match.

```
var intrinsic_strength = mem(function ({athlete}) {
   ... return Math.max(gaussian(intrinsic_strength_average,
   intrinsic_strength_variance), 0)})

var effort_level_in_match = mem(function({athlete, match}) {
     ...return Math.min(Math.max(gaussian(effort_average,
     effort_variance), 0), 1)})

var athlete_pulling_strength_in_match = mem(function({athlete, match}) {
     ...var pulling_strength_adjusted_for_effort_level_in_this_match =
      effort_multiplier * base_pulling_strength_in_action;
      return pulling_strength_adjusted_for_effort_level_in_this_match;})

var team_pulling_strength_in_match = function({team, match}) {
      return sum(map(function(athlete) {return
      athlete_pulling_strength_in_match({athlete: athlete,
      match: match}) }, team))}
```

Exp. 2
Excerpts from underspecified background *B*, **canoe racing**

In this event, the athletes are competing in a series of **canoe races**. In each race, the team that wins depends on the **average speed** with which the athletes are able to row, based on their **intrinsic strength** modulated by other factors including **how much effort** they put in to that race.

Corresponding excerpts from synthesized $M_{\text{ad-hoc}}$
Excerpts from informal retrieved background knowledge B^+

```
Intrinsic strength is an underlying attribute of a given athlete
that varies somewhat widely from athlete to athlete... We'll
imagine that players tend to be weak, average, or strong, and
uniformly around that.

The amount of effort that an athlete puts into any given race is a
continuous parameter...This effort is somewhat contingent on an
athlete's intrinsic strength, because stronger athletes probably
tend to be more likely to put in extra effort...

The speed of an individual athlete is determined by their
intrinsic strength and effort level.

In each race, the team that wins depends on the average speed with
which the athletes are able to row.
```

```
var intrinsic_strength = mem(function ({athlete}) {
      var intrinsic_strength_average = categorical({ps: [weak_prior,
      average_prior, strong_prior], vs: [weak_mean,
      average_mean,strong_mean]});...)}

var effort_level_in_race = mem(function({athlete, race}) {
     var parameter_low_effort_prior =
          intrinsic_strength(athlete) > strong_threshold ? 0.05 :
          intrinsic_strength(athlete) < weak_threshold ?  0.8 : 0.2; ...})

var athlete_rowing_speed_in_race = mem(function({athlete, race}) {
     var strength_modulated_for_effort_level = (effort_level_in_race({athlete:
     athlete, race: race}) / average_effort) * intrinsic_strength({athlete :
     athlete}); return strength_modulated_for_effort_level; })

var team_rowing_speed_in_race = function({team, race}) {
      var average_team_speed = mean(map(function(athlete) { return
      athlete_rowing_speed_in_race({athlete: athlete, race: race}) }, team));
      return average_team_speed;}
```

Exp. 3
Participant-generated commentaries
Corresponding excerpts from synthesized $M_{\text{ad-hoc}}$
Retrieved B^+

"Quinn is a newcomer to canoe races and this is his first competition. He doesn't have nearly the hours on the water any single other person in the water today does"

```
A canoe racer who is inexperienced with canoe racing puts
less effort and is probably less strong.
```

```
var inexperienced_canoe_racer = function({athlete}) {
    return effort_level({athlete}) < effort_threshold &&
    intrinsic_strength({athlete}) < strength_threshold;}
```

"Ness took an energy drink and it just kicked in to give them extra energy for their matchs"

```
Taking an energy drink is an event that affects the
match in which the energy drink was taken.
```

```
var pulling_strength_adjusted_for_taking_energy_drink =
    took_energy_drink ({athlete})
     ? pulling_strength* 1.3 : pulling_strength;
```

"In the first match of tug-of-war Kay pulled a muscle in their shoulder that limited their pulling output going forward."

```
Reduced if they pulled a muscle in their shoulder in any
PREVIOUS match.
```

```
var pulling_strength_adjusted_for_pulled_muscle =
   any_previous_time_inclusive(function(prev_match) {
    return pulled_muscle_in_shoulder_in_match({athlete,
    prev_match)}, match) ? base_pulling_strength_in_match * 0.7 :
    base_pulling_strength_in_match;
```

Figure 4: Excerpts showing key parts of the **natural language inputs**, retrieved additional **informal background knowledge** B^+ as natural language describing proposed relevant latent variables, and resulting **formal ad-hoc models** $M_{\text{ad-hoc}}$ as synthesized probabilistic programs. **Exp. 1** (top) shows how detailed natural language descriptions (left) are grounded into stochastic function definitions (right). **Exp. 2** (center) shows how our pipeline also retrieves relevant variables in natural language (left, bottom) that are formalized into resulting synthesized programs (right). **Exp. 3** (bottom) shows how *additional participant-provided free-form natural language* can also formalized into ad-hoc model definitions (excerpted on the left).

Together these results suggest that in each of these experimental settings people reason in ways that are consistent with normative Bayesian inferences over some structured set of relevant variables – and that these variables can be retrieved automatically using our approach.

Key finding 2: People's reasoning is more similar to inference in structured probabilistic models than LM-only alternatives, especially when generalizing to arbitrary new details. We next compare people's inferences to alternate models and compare models to one another to probe the structural cross-similarity in probabilistic judgments across model classes. The blocks of correlations

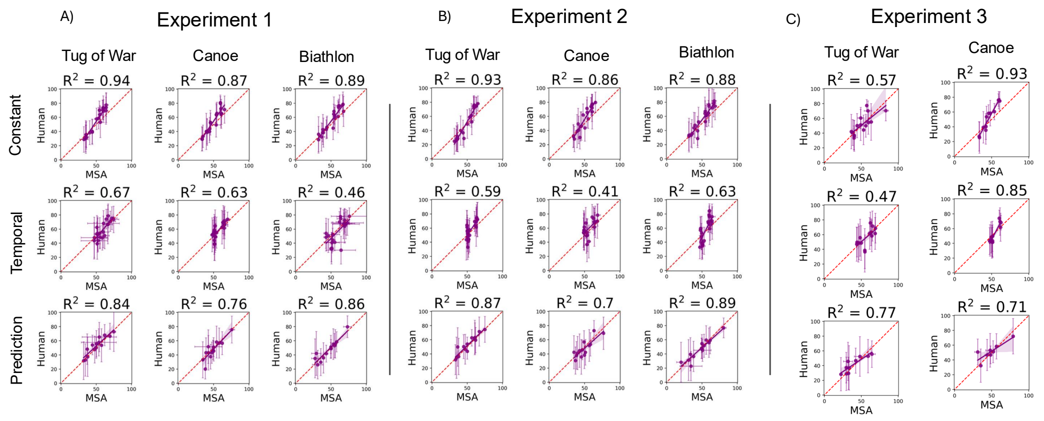


Figure 5: Correlations between human judgments and MSA predictions across Experiment 1 **(A)**, Experiment 2 **(B)**, and Experiment 3 **(C)**. Each plot shows correlations within a specific sport and query type. Points on each plot are between *mean* predictions for each query, for each scenario, over all participant answers and over all simulated participant model posteriors for that query.

visible on the heatmaps in Figure 6 (A-C) show that in many cases, humans are better correlated with themselves and symbolic models (MSAs and the hand-crafted probabilistic models) than with LMs, which are instead better correlated with each other. This finding parallels cross-model results from the distributional metric Figure 6 (D-F), which also shows that humans, MSA, and hand-crafted models distributions (blue, purple, and silver) are generally closer to the human distribution than the LM distribution (pink and orange).

In Experiment 3, which targeted the open-world setting, MSA judgments were substantially better aligned with people's than were those of LM-only alternatives. That these differences were greatest in this setting is suggestive in two ways. First, LM baselines may face particular challenges in fitting human judgments as the distribution shifts further from familiar settings. Second, the open-world Experiment 3 represents the most significant a priori challenge to existing, hand-crafted symbolic models of cognition. That the MSA continues to outperform LM baselines in this case suggests that probabilistic modeling can continue to best capture human judgments, even in the open-world setting. Globally, these findings suggest an asymmetry in LMs abilities — LM's may be relatively better equipped to retrieve relevant world knowledge in human-like ways, but relatively less able to integrate that evidence into a locally coherent world model the way that people do.

Key finding 3: MSA can retrieve and represent relevant information about arbitrary situations as structured probabilistic models. Both correlational and distributional analyses suggest that the implemented MSA can synthesize models that quantitatively capture human inferences. But what do these ad-hoc probabilistic models actually look like? The qualitative examples in Figure 4 show that the LM-guided synthesis approach is generally able to retrieve reasonable descriptions of variables and causal dependencies; and that it can parse natural language descriptions from both the original inputs (Figure 4, Exp. 1) and retrieved background knowledge (Figure 4, Exp. 2, 3) into corresponding probabilistic programs.

At the same time, manual inspection of the underlying code also reveals places where the MSA implementation generates imperfect parses. These range from somewhat minor (e.g., the retrieved additional natural language B^+ suggests that the winning team depends on a team's *average* speed, but the synthesized model sometimes does not encode this with a `mean`); to relatively more drastic omissions (for instance, the synthesized models in Exp. 3 often did not correctly interpret modal temporal logics, like that a pulled shoulder limits pulling strength in *future matches*, until we allowed the synthesis procedure to access a library of modal logic functions). We discuss both these limitations and opportunities for further work in Supplement Section A2.

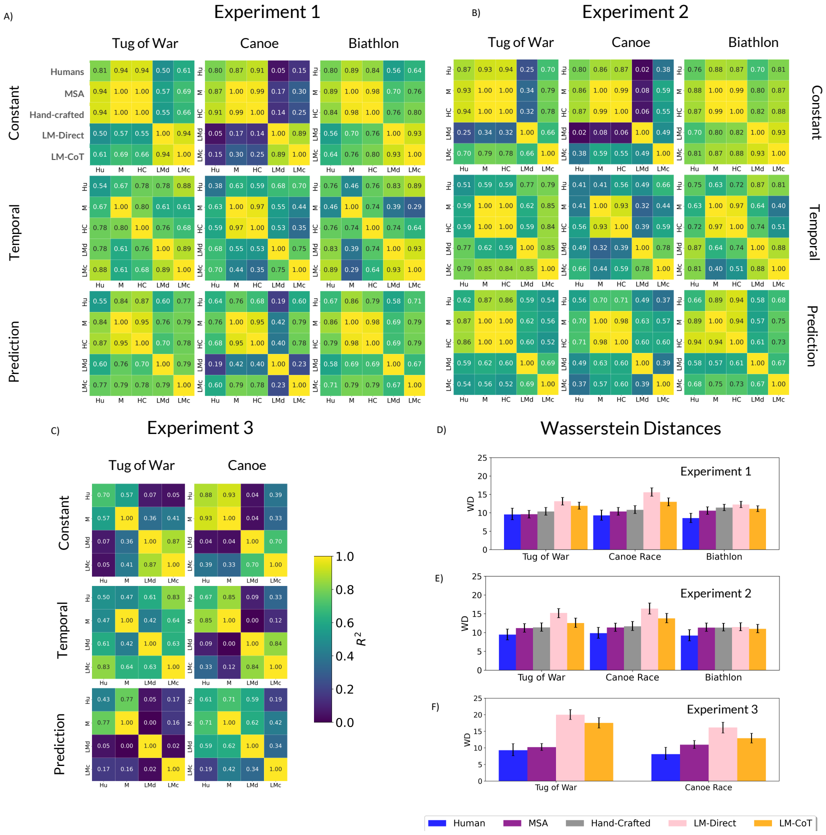


Figure 6: Correlational and distributional comparisons of MSA, hand-crafted symbolic, and LM-only alternative models to human judgments. **(A)-(C)** Cross-model and model-human R^2 for each experiment, where R^2 are computed over mean judgments per scenario per query. Reasoners are abbreviated (Hu=human, M=MSA, HC=hand-crafted symbolic probabilistic program, LMd=LM-Direct, LMc=LM-CoT); **(D)-(F)** Wasserstein Distances from models and baselines to distribution of human judgments (individual Wasserstein Distances (WDs) computed between judgments per query per scenario, then aggregated as the mean over query types, and mean across query types for each depicted sport and experiment; *lower* distances indicate closer similarity to the distribution of human judgments. Error bars show 95% CI over 1000 bootstrapped samples, with replacement.

5 Discussion

In this work, we investigated how people are able to reason in ways that deliver *global relevance* and *local coherence* – that is, how human reasoning is able to show both a sensitivity to relevant considerations from across people's background knowledge and coherent integration of evidence over those considerations. In our experiments, we found that an MSA can synthesize ad-hoc models that fit human judgments in the first instance, and fit those judgments better than LM baselines. This suggests that MSAs offer a promising avenue towards capturing the computations underlying human reasoning, especially in open-world settings. We include an expanded discussion of limitations and future work in Supplement A2. The problem of open-world cognition is the challenge of being able to reason well-enough in the vast space of problems we encounter. We have taken a small step in that direction by showing that reasonable mental models can be automatically synthesized for new problem instances in a novel family of tasks. Much more is needed to determine whether this approach can scale to the level of generality and flexibility seen in human cognition.

References

[1] Webppl probabilistic programming language. URL https://github.com/probmods/webppl.

[2] J. Andreas. Language models, world models, and human model-building. *Language & Intelligence@ MIT*, 2024.

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

[4] P. W. Battaglia, J. B. Hamrick, and J. B. Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110(45):18327–18332, 2013. doi: 10.1073/pnas.1306572110. URL https://www.pnas.org/doi/pdf/10.1073/pnas.1306572110.

[5] E. Bingham, J. P. Chen, M. Jankowiak, F. Obermeyer, N. Pradhan, T. Karaletsos, R. Singh, P. Szerlip, P. Horsfall, and N. D. Goodman. Pyro: Deep universal probabilistic programming. *Journal of machine learning research*, 20(28):1–6, 2019.

[6] M. Binz, E. Akata, M. Bethge, F. Brändle, F. Callaway, J. Coda-Forno, P. Dayan, C. Demircan, M. K. Eckstein, N. Éltető, et al. Centaur: a foundation model of human cognition. *arXiv preprint arXiv:2410.20268*, 2024.

[7] T. Brooke-Wilson. *Bounded Rationality as a Strategy for Cognitive Science*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, 2023. URL https://philosophy.mit.edu/wp-content/uploads/brookewilson_dissertation.pdf.

[8] B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76:1–32, 2017.

[9] P. S. Castro, N. Tomasev, A. Anand, N. Sharma, R. Mohanta, A. Dev, K. Perlin, S. Jain, K. Levin, N. Éltető, et al. Discovering symbolic cognitive models from human and animal behavior. *bioRxiv*, pages 2025–02, 2025.

[10] K. Chandra, J. Ragan-Kelley, and J. Tenenbaum. Theories of mind as languages of thought for thought about thought. 2025.

[11] K. M. Collins, I. Sucholutsky, U. Bhatt, K. Chandra, L. Wong, M. Lee, C. E. Zhang, T. Zhi-Xuan, M. Ho, V. Mansinghka, et al. Building machines that learn and think with people. *Nature Human Behavior*, 2024.

[12] K. J. W. Craik. *The nature of explanation*, volume 445. CUP Archive, 1943.

[13] L. Cross, V. Xiang, A. Bhatia, D. L. Yamins, and N. Haber. Hypothetical minds: Scaffolding theory of mind for multi-agent tasks with large language models. *arXiv preprint arXiv:2407.07086*, 2024.

[14] M. F. Cusumano-Towner, F. A. Saad, A. K. Lew, and V. K. Mansinghka. Gen: a general-purpose probabilistic programming system with programmable inference. In *Proceedings of the 40th acm sigplan conference on programming language design and implementation*, pages 221–236, 2019.

[15] D. Dohan, W. Xu, A. Lewkowycz, J. Austin, D. Bieber, R. G. Lopes, Y. Wu, H. Michalewski, R. A. Saurous, J. Sohl-Dickstein, et al. Language model cascades. *arXiv preprint arXiv:2207.10342*, 2022.

[16] J. Domke. Large language bayes. *arXiv preprint arXiv:2504.14025*, 2025.

[17] K. Ellis, C. Wong, M. Nye, M. Sablé-Meyer, L. Morales, L. Hewitt, L. Cary, A. Solar-Lezama, and J. B. Tenenbaum. Dreamcoder: Bootstrapping inductive program synthesis with wake-sleep library learning. In *Proceedings of the 42nd acm sigplan international conference on programming language design and implementation*, pages 835–850, 2021.

[18] Y. Feng, B. Zhou, W. Lin, and D. Roth. Bird: A trustworthy bayesian inference framework for large language models. *arXiv preprint arXiv:2404.12494*, 2024.

[19] D. Gentner and D. R. Gentner. Flowing waters or teeming crowds: Mental models of electricity. In D. Gentner and A. L. Stevens, editors, *Mental Models*, pages 99–129. Lawrence Erlbaum Associates, Hillsdale, NJ, 1983.

[20] T. Gerstenberg and J. B. Tenenbaum. Intuitive theories. In *Oxford handbook of causal reasoning*, pages 515–548. 2017.

[21] N. Goodman, V. Mansinghka, D. M. Roy, K. Bonawitz, and J. B. Tenenbaum. Church: a language for generative models. *arXiv preprint arXiv:1206.3255*, 2012.

[22] N. D. Goodman, J. B. Tenenbaum, and T. Gerstenberg. Concepts in a probabilistic language of thought. In E. Margolis and S. Laurence, editors, *The conceptual mind: New directions in the study of concepts*, pages 59–109. MIT Press, Cambridge, MA, 2014.

[23] A. Gopnik, C. Glymour, D. M. Sobel, L. E. Schulz, T. Kushnir, and D. Danks. A theory of causal learning in children: causal maps and bayes nets. *Psychological review*, 111(1):3, 2004.

[24] G. Grand, L. Wong, M. Bowers, T. X. Olausson, M. Liu, J. B. Tenenbaum, and J. Andreas. Lilo: Learning interpretable libraries by compressing and documenting code. In *The Twelfth International Conference on Learning Representations*, 2024.

[25] G. Grand, J. B. Tenenbaum, V. K. Mansinghka, A. K. Lew, and J. Andreas. Self-steering language models. *arXiv preprint arXiv:2504.07081*, 2025.

[26] J. B. Hamrick, K. A. Smith, T. L. Griffiths, and E. Vul. Think again? the amount of mental simulation tracks uncertainty in the outcome. In *Proceedings of the annual meeting of the cognitive science society*, volume 37, 2015.

[27] M. K. Ho, D. Abel, C. G. Correa, M. L. Littman, J. D. Cohen, and T. L. Griffiths. People construct simplified mental representations to plan. *Nature*, 606(7912):129–136, 2022.

[28] Z. Hu and T. Shu. Language models, agent models, and world models: The law for machine reasoning and planning. *arXiv preprint arXiv:2312.05230*, 2023.

[29] T. Icard and N. D. Goodman. A resource-rational approach to the causal frame problem. In *CogSci*, 2015.

[30] J. Jara-Ettinger, H. Gweon, L. E. Schulz, and J. B. Tenenbaum. The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences*, 20(8):589–604, 2016. doi: 10.1016/j.tics.2016.05.011. URL https://www.sciencedirect.com/science/article/pii/S1364661316300535.

[31] P. N. Johnson-Laird. Mental models in cognitive science. *Cognitive science*, 4(1):71–115, 1980.

[32] A. K. Lew, M. H. Tessler, V. K. Mansinghka, and J. B. Tenenbaum. Leveraging unstructured statistical knowledge in aprobabilistic language of thought. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 42, 2020.

[33] M. Y. Li, E. B. Fox, and N. D. Goodman. Automated statistical model discovery with language models. *arXiv preprint arXiv:2402.17879*, 2024.

[34] F. Lieder and T. L. Griffiths. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, 43:e1, 2020.

[35] J. Loula, B. LeBrun, L. Du, B. Lipkin, C. Pasti, G. Grand, T. Liu, Y. Emara, M. Freedman, J. Eisner, et al. Syntactic and semantic control of large language models via sequential monte carlo. *arXiv preprint arXiv:2504.13139*, 2025.

[36] R. T. McCoy, S. Yao, D. Friedman, M. Hardy, and T. L. Griffiths. Embers of autoregression: Understanding large language models through the problem they are trained to solve. *arXiv preprint arXiv:2309.13638*, 2023.

[37] I. Mirzadeh, K. Alizadeh, H. Shahrokhi, O. Tuzel, S. Bengio, and M. Farajtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229*, 2024.

[38] W. T. Piriyakulkij, Y. Liang, H. Tang, A. Weller, M. Kryven, and K. Ellis. Poe-world: Compositional world modeling with products of programmatic experts. *arXiv preprint arXiv:2505.10819*, 2025.

[39] J. Richens, D. Abel, A. Bellot, and T. Everitt. General agents need world models. *arXiv preprint arXiv:2506.01622*, 2025.

[40] M. Rmus, A. K. Jagadish, M. Mathony, T. Ludwig, and E. Schulz. Generating computational cognitive models using large language models, 2025. URL https://arxiv.org/abs/2502.00879.

[41] H. Tang, D. Key, and K. Ellis. Worldcoder, a model-based llm agent: Building world models by writing code and interacting with the environment. *Advances in Neural Information Processing Systems*, 37:70148–70212, 2024.

[42] K. Vafa, J. Y. Chen, J. Kleinberg, S. Mullainathan, and A. Rambachan. Evaluating the world model implicit in a generative model. *arXiv preprint arXiv:2406.03689*, 2024. URL https://arxiv.org/abs/2406.03689.

[43] K. Valmeekam, M. Marquez, A. Olmo, S. Sreedharan, and S. Kambhampati. Planbench: An extensible benchmark for evaluating large language models on planning and reasoning about change. *Advances in Neural Information Processing Systems*, 36:38975–38987, 2023.

[44] G. Wang, Y. Xie, Y. Jiang, A. Mandlekar, C. Xiao, Y. Zhu, L. Fan, and A. Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023.

[45] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[46] L. Wong, G. Grand, A. K. Lew, N. D. Goodman, V. K. Mansinghka, J. Andreas, and J. B. Tenenbaum. From word models to world models: Translating from natural language to the probabilistic language of thought. *arXiv preprint arXiv:2306.12672*, 2023.

[47] L. Wong, J. Mao, P. Sharma, Z. Siegel, J. Feng, N. Korneev, J. B. Tenenbaum, and J. Andreas. Learning adaptive planning representations with natural language guidance. International Conference on Learning Representations, 2024.

[48] S. Xia, B. Lu, and J. Eisner. Let's think var-by-var: Large language models enable ad hoc probabilistic reasoning. *arXiv preprint arXiv:2412.02081*, 2024.

[49] L. Ying, K. M. Collins, M. Wei, C. E. Zhang, T. Zhi-Xuan, A. Weller, J. B. Tenenbaum, and L. Wong. The neuro-symbolic inverse planning engine (nipe): Modeling probabilistic social inferences from linguistic inputs. *arXiv preprint arXiv:2306.14325*, 2023.

[50] L. Ying, R. Truong, K. M. Collins, C. E. Zhang, M. Wei, T. Brooke-Wilson, T. Zhi-Xuan, L. Wong, and J. B. Tenenbaum. Language-informed synthesis of rational agent models for grounded theory-of-mind reasoning on-the-fly, 2025. URL https://arxiv.org/abs/2506.16755.

[51] L. Ying, T. Zhi-Xuan, L. Wong, V. Mansinghka, and J. B. Tenenbaum. Understanding epistemic language with a language-augmented bayesian theory of mind. *Transactions of the Association for Computational Linguistics*, 2025.

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

[53] Z. Zhang, C. Jin, M. Y. Jia, and T. Shu. Autotom: Automated bayesian inverse planning and model discovery for open-ended theory of mind. *arXiv e-prints*, pages arXiv–2502, 2025.

[54] T. Zhi-Xuan. *Pddl. jl: An extensible interpreter and compiler interface for fast and flexible ai planning*. PhD thesis, Massachusetts Institute of Technology, 2022.