Grounded physical language understanding with probabilistic programs and simulated worlds

Cedegao E. Zhang[1] **Lionel Wong**[1] **Gabriel Grand**[2] **Joshua B. Tenenbaum**[1,2]
[1]BCS, MIT [2]CSAIL, MIT
{cedzhang, zyzzyva, grandg, jbt}@mit.edu

Abstract

Human language richly invokes our intuitive physical knowledge. We talk about physical objects, scenes, properties, and events; and we can ask questions and answer them with predictions and inferences about physical worlds described entirely in language. How does language construct meanings that connect to our general physical reasoning? In this paper, we propose **PiLoT**, a computational model that maps language into a *probabilistic language of thought*—meanings are constructed as probabilistic programs, which provide a formal basis for probabilistic and physical reasoning. Our model uses a large language model (LLM) to map from language to meanings and a probabilistic physics engine to support inferences over scenes described in language. We conduct a linguistic reasoning experiment based on prior psychophysics studies that requires reasoning about physical outcomes based on linguistic descriptions. We show that PiLoT well predicts human judgments across this experiment and outperforms baseline models which use the LLM to directly perform the same task.

Keywords: natural language understanding; probabilistic programming; world model; semantic parsing; intuitive physics

Introduction

Physical intuitions pervade our everyday language. We can describe and imagine a *tall stack of plates*, a *heavy box*, and objects that *fall*, *bounce*, and *collide*. We flexibly answer questions that require physical prediction (*what will happen if a kid crashes into that table stacked with plates?*) or inference (*how heavy is that box that no one can lift?*). Our intuitions hold even when language is vague (*how tall is tall?*) or we are uncertain about aspects of the world itself.

How does the meaning we make from *language* drive this kind of physical reasoning? That is, how does language construct the mental representations that allow us to imagine these possible worlds or answer questions about them based on our physical knowledge? The *formal semantics* tradition emphasizes the importance of compositionality and considers truth conditions central to meaning, but it does not directly engages with how meaning connects to cognitive mechanisms (Heim & Kratzer, 1998). One influential cognitive account of semantics suggests that meanings are *simulators*, such that language constructs composable representations for mental simulation (Barsalou et al., 2008). Other work has considered specifically how structured representations of meaning convey physical information, with a focus on how verb meanings may be realized into cognitively-grounded physical concepts of motion and forces (Talmy, 1988; Levin, 1993; Schuler, 2005). Finally, *distributional semantics* accounts suggest that certain aspects of linguistic meaning are correlated with the statistical distribution of words used in context (Harris, 1954; Chater et al., 2006), which may include latent information about the physical world.

These accounts leave open important questions for a complete computational account bridging language and physical reasoning. What formal representations of meaning can support and drive mental simulation, allowing us to tractably imagine and run simulations over arbitrary scenes described in language? How can language abstractly convey many possible worlds—such as the many worlds in which *there are some plates in a tall stack on a table*—so that we can still ground these worlds in simulation or physical knowledge?

Further, how do the representations of linguistic meaning relate to those that allow us to reason about physics *independent* of language? Extensive developmental evidence suggests that, even prior to acquiring language, we have a core understanding of the physical principles that govern our world (Spelke, 1990; Spelke et al., 1995; Baillargeon, 2004; Hespos & Baillargeon, 2008; Rips & Hespos, 2015). A productive line of computational cognitive models, in turn, has modeled human physical understanding as probabilistic inference over a *mental physics engine*, using representations like those for simulations in video games (T. D. Ullman et al., 2010; T. Ullman et al., 2012; Battaglia et al., 2013; T. D. Ullman et al., 2017). But how are these capabilities integrated with language, allowing us to imagine and draw inferences over possible physical worlds that we describe in words?

In this paper, we propose **PiLoT** (*Physics in a Language of Thought*), a computational model that **maps language into a *probabilistic language of thought* which supports physical simulation and inference** (Fig. 1). We propose this as a modular, cognitive framework that relates meaning in language to general physical reasoning abilities from broader cognition. We aim to unify and formalize distinct aspects of meaning in this domain that we have discussed—meanings as *compositional and symbolic representations*, meanings as *representations for mental and physical simulation*, and meanings as correlated with the *distributional statistics of language*—within an overarching computational model.

Our model builds directly on a theoretical background that suggests we construct linguistic meaning from cognitive representations in a compositional *language of thought* (Fodor, 1975; Jackendoff, 1985; Lakoff, 1988) and more recent pro-

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

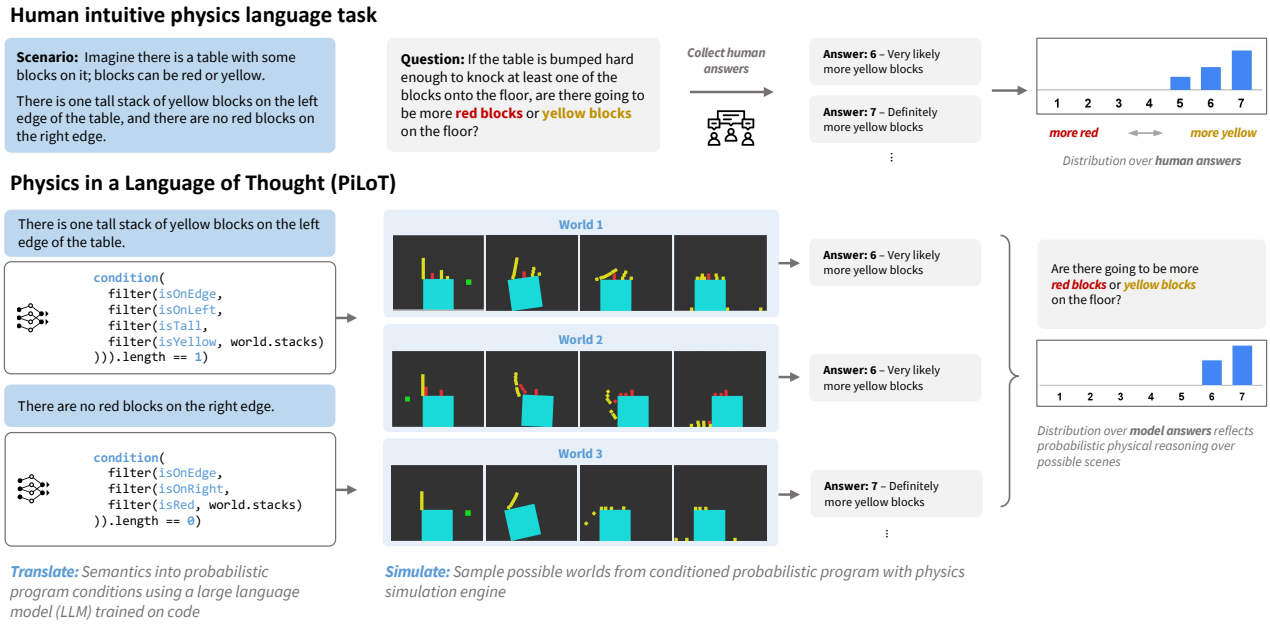

Figure 1: Human language understanding draws on our flexible, intuitive physical knowledge. (*Top*) We measure human judgements on a domain of linguistic scene reasoning tasks (inspired by Battaglia et al. (2013)), asking subjects about physical outcomes based on descriptions of a tabletop scene with varying configurations of red and yellow blocks. (*Bottom*) Our model, **PiLoT**, reasons about these descriptions by first translating language into formal program semantics using a large statistical *language-code translation model*. Our semantics are *probabilistic program expressions* that condition a generative model over possible scenes. To answer questions about physical outcomes, our model then samples and simulates scenes from the conditioned model using a *physics engine*, producing probabilistic inferences that correlate well with human judgments.

posals that address uncertainty in meaning using probabilistic semantic representations (van Eijck & Lappin, 2012; Cooper et al., 2015; Goodman & Lassiter, 2015). By modeling semantics as *probabilistic programs grounded in a physics engine*, we show how language can map into symbolic meaning representations that in turn support probabilistic, physical simulation and physical inferences over language.

One open question for many symbolic linguistic theories, however, has long been how we might actually implement a broad-coverage and context-sensitive meaning function. In this work, we propose that the meaning functions can be instantiated as general joint distributions between natural language and symbolic representations. We model this using *large language models* (LLMs) trained on code to translate between sentences in language and symbolic semantic expressions and show that this approach can generalize across a broad range of sentences in context. LLMs are clearly not trained in cognitively plausible ways, but we use them here as an in-principle instantiation of the distributional semantics hypothesis, suggesting how it can be used to *relate* language and symbolic representations learned from prior joint distributions over both.

This model draws on and extends important ideas from the most closely related computational work, towards a more complete account of robust, human-like physical reasoning over language. As with Liu et al. (2022), we show that LLMs can robustly parse language into programs, and interface with a physics engine to ground language in physical simulation. To extend this approach towards the probabilistic judgments we make about everyday physical language, we build on prior computational cognitive approaches in linguisitics and psychophysics. Inspired by Goodman & Lassiter (2015), we parse language into a *probabilistic* programming language, formalizing how meanings can condition and support inferences over possible worlds. By integrating these semantics with the probabilistic physics engine used in Battaglia et al. (2013), we show how language can support flexible, probabilistic inferences about the physical world.

We evaluate this model in comparison to human behavior using a new domain of *linguistic physical reasoning tasks*. These tasks are inspired by an existing battery of visual psychophysics tasks (Téglás et al., 2011; Battaglia et al., 2013), but designed to evaluate how these prior results on physical reasoning relate to information conveyed linguistically. Our base task combines descriptions of physical scenes of varying object configurations (*Imagine a table with some red and yellow blocks on it*) with a simple but general scene understanding challenge (*if the table is bumped, will there be more red or yellow blocks knocked to the floor?*). We construct a dataset of stimuli spanning a range of linguistic constructions, and evaluate human judgments on our dataset.

We show that PiLoT **robustly predicts human reasoning about these linguistic physical scenes**. Our model also better correlates with human judgements than an ablated version

Easy (1 concept)	*There are four stacks of red blocks, and there is one stack of yellow blocks.* **[Numbers]**	# 1
	There are short stacks of red blocks, and there are short stacks of yellow blocks. **[Graded adjectives]**	# 16
Moderate (2 concepts)	*There are many yellow blocks on the left side of the table, there are no blocks on the middle, and there are no blocks on the right side.* **[Spatial relations, quantifiers]**	# 29
	There are stacks of yellow blocks, and there are stacks of red blocks. All of the yellow stacks are tall, and all of the red stacks are short. **[Quantifiers, graded adjectives]**	# 38
Challenging (3-4 concepts)	*There is one stack of yellow blocks on the center of the table, and there is one tall stack of red blocks near the yellow stack.* **[Numbers, spatial relations, graded adjectives]**	# 49
	There are at least five stacks of blocks on the table. No more than half of the stacks are tall. Most of the stacks are red, and most of the stacks are on the right side. **[Numbers, spatial relations, quantifiers, graded adjectives]**	# 64

Table 1: Example stimuli from our linguistic physical reasoning experiment, describing configurations of blocks on a table. Scene descriptions are parameterized based on distinct conceptual categories, and vary in complexity based on how many distinct conceptual kinds are invoked in a given description.

of our own model, in which we directly query a large statistical language model to predict physical inferences on these same tasks. We also find that our model **robustly predicts the underlying distribution of human judgments**, capturing the uncertainty inherent to how we reason about abstract, linguistic descriptions about these scenes.

Linguistic physical reasoning experiment

We begin by describing the human experiment and domain that we use, to provide intuition for the model used in the remainder of this paper. Our linguistic and physical reasoning task was inspired by psychophysics stimuli from Téglás et al. (2011) and Battaglia et al. (2013), in which subjects were presented with visual scenes involving different configurations of red and yellow blocks stacked on a table and asked to predict physical outcomes. Our linguistic stimuli adapts this domain to scenes described in *language*. Unlike visual images, this task requires reasoning over the additional uncertainty inherent to language – a sentence like *There are three stacks of red blocks on the table* leaves open where these stacks might be located, or how tall they might be.

Each stimuli in our experiment begins with a linguistic description of the general domain of scenes (*Imagine a table with some red or yellow blocks on it*), then provides varying additional information about the block configuration (*There are at least two tall stacks of yellow blocks on the right edge of the table*). Based on each scene description, we pose a simple linguistic query that requires reasoning about possible physical outcomes: *If the table is bumped hard enough to knock at least one of the blocks onto the floor, are there going to be more red blocks or yellow blocks on the floor?* Future work can easily adapt these stimuli to other queries, such as specifying the direction or magnitude of the bump.

Using this base template, we design *64 scene reasoning stimuli* that vary systematically over a space of linguistic concepts, and in the complexity of each scene description. Scene descriptions were parameterized based on the following conceptual categories, each widely studied in both cognitive science and natural language semantics:

- **Number**: how many blocks or stacks are on a table, such as *three* stacks of red blocks or *two* yellow blocks (Bartsch, 1973; Gelman & Gallistel, 1986; Carey, 2009).
- **Spatial relations**: prepositions describing where blocks are located, such as the *center*, *left and right sides*, and *left and right edges* of the table, or *near* another block on the table (Landau & Jackendoff, 1993).
- **Quantifiers**: quantifiers such as *many*, *few*, *several*, *most*, or *half* of the blocks being of a certain color, position, etc., and negations such as *none* of the blocks being a certain color, etc. (Barwise & Cooper, 1981; van Tiel et al., 2021).
- **Graded adjectives**: adjectives describing the stacks as *tall*, *very tall*, *short*, etc. (Klein, 1980; Williamson, 2002).

Using these base concepts, we vary stimuli complexity based on how many distinct classes of concepts are invoked in a given scene description. Our experiment comprises 16 **easy** stimuli, which contain concepts from a single conceptual category; 24 **moderate** complexity stimuli, containing concepts from two categories; and 24 **challenging** stimuli, which contain concepts from 3-4 categories (examples in Table 1).

For the experiment, we collect and evaluate human judgments on these linguistic scene reasoning tasks. Subjects produced judgments about each stimulus on a 1–7 Likert scale of confidence spanning *1 (definitely more red blocks)* to *7 (definitely more yellow blocks)*, measuring subject uncertainty about an inherently probabilistic task.

In total, we recruit 160 from Prolific; each viewed a random batch of 16 stimuli. We collect approximately 40 human responses per stimulus. Participants were native English speakers from the USA/UK and received payments at $15/hr.

Our model: PiLoT

Our model relates mental physical simulation with a broad-coverage mapping function from language into a symbolic language of thought. This model, PiLoT, consists of three modules: a probabilistic generative model over possible scenes, a language-to-code translation model, and a physics simulator. Together, the generative model and physics simulator implement a version of the model used in Battaglia et al. (2013). The translation model extends this framework to show how it can generally integrate natural language, in the spirit of Goodman & Lassiter (2015).[1]

[1]The code excerpts presented in this section have been simplified for legibility. The full code of the model is available at https://tinyurl.com/phys-lang.

Probabilistic generative model We begin by defining a generative model over possible worlds in our blockworld domain. We write this model in WebPPL (Goodman & Stuhlmüller, 2014), a probabilistic programming language based on JavaScript. For instance, to construct a new block stack, the model makes a series of random choices to determine the stack's color, height, and position on the table:

```
var blockColor = function () {
  return flip() ? 'red' : 'yellow'
}
var stackHeight = function () {
  return geometric(0.7, 1, 8)
}
var xPositionOnTable = function (table) {
  return uniformDraw(
    _.range((worldWidth / 2) - table.width,
            (worldWidth / 2) + table.width))}
var newStack = {
  color: blockColor(),
  height: stackHeight(),
  x: xPositionOnTable(table), }
```

The stochasticity that arises from these random choices is what makes our model *probabilistic*. Each call to `makeBlockWorld()` (below) returns a different blockworld with a variable number of stacks (between 1 and 8) in different configurations. Thus, `makeBlockWorld()` defines a probability distribution over possible worlds and running it produces a sample from an uninformed prior.

```
var makeBlockWorld = function () {
  var stacks = buildStacks(numStacks)
  var world = {
    stacks: stacks,
    blocks: getBlockList(stacks),
    table: { shape: 'rect', dims: [tableSize,
    ↪  tableSize], x: worldWidth / 2, ... },
    force: generateForce(velocity, direction),
  }
  return world }
```

Additionally, our model includes a set of functions that collectively define a *domain semantics*. By composing statements in the semantics, we can model the meanings of various linguistic utterances. As a simple example:

```
var isRed = function (obj) {
  return obj.color == 'red'
}
var isTall = function (stack) {
  return stack.height >= th_Tall
}
var isOnLeft = function (obj) {
  return obj.x <= th_Left
}
var isNear = function (obj1) {
  return function (obj2) {
    return abs(obj1.x - obj2.x) <= th_Near}}

// There is a tall stack of red blocks on the left
↪  side of the table.
condition(filter(isTall, filter(isRed,
↪  filter(isOnLeft, world.blocks)).length == 1))
```

In WebPPL, calling `condition()` constrains samples from the generative model to be consistent with the conditioning statement. In the above example, the conditioned model returns only blockworlds that have a tall stack of red blocks on the left side of the table. Condition statements deliberately admit imprecision (e.g., "There are at least two red blocks...") and can be added sequentially as new information is available. In this way, conditioning provides a natural way to model a reasoner with some prior over scenes who incrementally updates their beliefs to form a posterior over possible worlds.

Language-to-code translation model Given a model of the world expressed in a PPL, we can frame the problem of language understanding as *language-to-code translation*. In this work, we focus on the subproblem of translating linguistic utterances about the state of a blockworld into conditioning statements that capture the semantics of the language. However, since the generative domain theory and the query are themselves WebPPL code, the same methods we use here could be adapted to translate these as well.

For our translation model, we leveraged the few-shot prompting capabilities of OpenAI's `Codex` model (Chen et al., 2021), a GPT language model fine-tuned on publicly available code from GitHub. For each task, we automatically constructed a prompt by concatenating the generative model code and 10 randomly-sampled examples from our domain, eachs manually annotated with code translations. Since JavaScript is prevalent in this corpus, we found `Codex` to be a highly adept translator for our domain, requiring little prompt engineering to produce robust translations of non-trivial phrases. For instance, in our experiments, Codex correctly translated the following from the "Challenging" category:

```
// There are two tall stacks of yellow blocks near
↪  the red stack.
condition(filter(isNear(filter(isOnEdge,
↪  filter(isRed, world.stacks))[0]), filter(isTall,
↪  filter(isYellow, world.stacks))).length == 2)

// Less than half of the blocks on the edges are
↪  yellow.
condition(filter(isYellow, filter(isOnEdge,
↪  world.blocks)).length < filter(isOnEdge,
↪  world.blocks).length / 2)
```

Queries to `Codex` were issued via the OpenAI API with temperature = 0 to ensure that translations adhered to domain semantics, and facilitate reproducibility.

Physics simulator To model tasks in our experiment, we interface our model with a *physics simulator* provided by the Box2D game engine (Catto, 2023). To simulate the table being bumped, we initialize each world with a high-velocity, bullet-like object that collides with the table. By randomly sampling and simulating multiple such worlds, we can obtain a distribution over outcomes. In this case, we are interested in the relative number of red and yellow blocks on the ground, which we normalize to a 7-point Likert scale, as below.

```
var simulateWorld = function () {
  var results = Infer(
    {method: 'rejection', samples: 10},
```

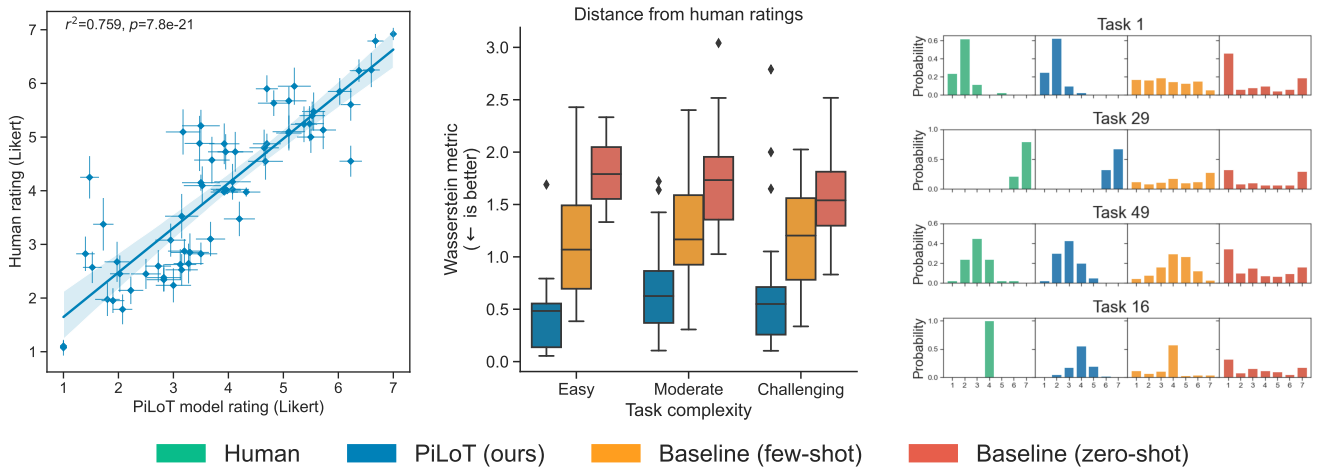

Figure 2: Comparison of PiLoT and baseline models to human ratings at increasing levels of granularity. Left: PiLoT broadly correlates with human Likert ratings across the 64 tasks in our experiment. (Vertical and horizontal bars reflect standard error for humans and PiLoT, respectively.) Middle: At each task complexity, PiLoT achieves closer fidelity to human ratings than the two baselines, as measured by Wasserstein distance. Right: Across individual tasks, humans (green) modulate their predictions to reflect differences in the scenarios. PiLoT generally mirrors human ratings distributions (top three rows), while the zero-shot baseline tends to be bimodal. (See Table 1 for the descriptions associated with each task.)

	Overall		**Easy**		**Moderate**		**Challenging**	
	R^2	WD	R^2	WD	R^2	WD	R^2	WD
Baseline (zero-shot)	0.40^{***}	1.69 (0.05)	0.73^{***}	1.82 (0.08)	0.37^{**}	1.75 (0.10)	0.16 (*N.S.*)	1.55 (0.09)
Baseline (few-shot)	0.34^{***}	1.20 (0.06)	0.54^{**}	1.17 (0.15)	0.43^{***}	1.22 (0.10)	0.06 (*N.S.*)	1.19 (0.10)
PiLoT (ours)	$\mathbf{0.76}^{***}$	**0.62** (0.07)	$\mathbf{0.91}^{***}$	**0.45** (0.10)	$\mathbf{0.78}^{***}$	**0.69** (0.09)	$\mathbf{0.69}^{***}$	**0.67** (0.13)
	Number		**Spatial**		**Quant. / Neg.**		**Graded Adj.**	
Baseline (zero-shot)	0.27^{**}	1.63 (0.06)	0.23^{**}	1.67 (0.08)	0.47^{***}	1.70 (0.08)	0.23^{*}	1.63 (0.08)
Baseline (few-shot)	0.15^{*}	1.19 (0.07)	0.17^{*}	1.21 (0.08)	0.36^{***}	1.28 (0.08)	0.30^{**}	1.14 (0.09)
PiLoT (ours)	$\mathbf{0.76}^{***}$	**0.57** (0.08)	$\mathbf{0.67}^{***}$	**0.74** (0.10)	$\mathbf{0.76}^{***}$	**0.67** (0.10)	$\mathbf{0.80}^{***}$	**0.54** (0.07)

Table 2: Performance of PiLoT and baseline models in comparison to humans, showing Pearson's R^2 and Wasserstein distance (WD) from human ratings. Top half: Results segmented by task complexity. Bottom half: Results segmented by conceptual category. P-value thresholds: $* = P < 0.05$, $** = P < 0.01$, $*** = P < 0.001$, N.S. = not significant.

```
    function () { return run(makeBlockWorld()) }
)
var pRed = exp(results.score('moreRed'))
var pYellow = exp(results.score('moreYellow'))
return round((pYellow / (pRed + pYellow)) * 6) + 1}
```

Intuitively, setting `samples` to 10 mirrors prior findings that humans perform 5-13 mental simulations when answering questions about similar block worlds (Battaglia et al., 2013). In our experiments, we run `simulateWorld()` 40 times for each task, effectively simulating 40 participants who each produce a Likert rating based on 10 mental simulations.[2]

Model experiments

To compare human and model performance, we conduct an analogous experiment using our linguistic reasoning tasks with our model and two baseline language models.

Our model To directly compare our model with human performance, our experiment simulates model answers to each stimulus on the same discretized 1-7 scale. For each stimulus, we translate the linguistic scene description into condition statements, sample and simulate $n = 10$ scenes from the conditioned generative program, and construct a sample-based estimate over the distribution of scenes in which more blocks of a given color fall to the floor. For each stimulus, we then simulate $n = 40$ independent sample-based inferences.

LLM-only baselines We also conduct two baseline experiments using the same distributional language model (Codex) to directly provide probabilistic judgments about each scene description, with no program semantics or physics engine.

- **Zero-shot LLM**: This baseline directly prompts an LLM with only the exact linguistic setup provided to human subjects in the human experiment (*In this experiment, you will read descriptions of scenes...*), followed by each individual stimulus (eg. *There are two red blocks on the table*) and the

[2]We note that, while the physics simulation has various hyperparameters, it offers robust out-of-box performance; indeed, manually tuning the hyperparameters to directly optimize for performance on our tasks yielded marginal improvements of $R^2 < 0.04$ relative to the naive settings that were used in our experiments.

question. We measure model responses over the same 1-7 scale of confidence by calculating normalized token log-probabilities for each scale item shown to humans.

- **Few-shot LLM**: This baseline augments the LLM query with a set of in-context examples of correct task/answer pairs (Brown et al., 2020). Prior to querying the model with a given stimulus, we additionally prompt the model with $n = 10$ (*stimulus*, *human response*) examples randomly sampled from heldout stimuli and the human responses.

Results and discussion

We first evaluate our model and baselines in comparison to human performance across tasks in the linguistic reasoning experiment (Table 2, *Overall*). We find that:

Our model best predicts human judgments across the physical language experiment. We calculate correlations between human judgments and our model based on mean per-stimulus judgments across human subjects, and across simulated Likert-scale judgments, and find that our model is significantly correlated with human judgements in the experiment overall (Fig. 2, $R^2 = 0.759$, $p < 0.001$). We calculate correlations between mean human judgments and a weighted mean per-stimulus judgment from the probability mass that the LLMs assign to each 1-7 scale value. Table 2 (Overall, R^2) shows that our model greatly outperforms both baselines in predicting human judgments.

Our model best captures the *distribution* of human judgments on each stimulus. We also calculate Wasserstein Distances between the human distribution of judgments predicted for each stimulus, and the distribution of judgments from our model and both baselines. Table 2 (*WD*) shows that our model also is much closer to the distribution of human judgments than either baseline. Qualitative inspection (Fig. 2) shows more revealing trends. The zeroshot model often produces contradictory, extreme judgments (1 or 7); and the fewshot model is often relatively uniform.

Next, we consider how stimuli complexity and specific conceptual categories impact model performance. We find:

Our model is much more robust as stimuli increase in complexity. Table 2 (*Easy, Moderate, Challenging*) shows that all models (ours, and both baselines) grow worse at predicting human behavior as stimuli complexity increases. However, our model is far more robust to stimuli complexity; the baselines grow rapidly less correlated with human judgments as complexity increases, and on the most challenging stimuli, our model still well-predicts human judgments ($R^2 = 0.69$, $p < 0.001$), whereas neither baseline is significantly correlated with human behavior.

LLM baselines struggle with number and spatial relations Table 2 (bottom half) also suggests that LLM baselines perform unevenly across the varying kinds of concepts in these stimuli. Both baselines appear strongest within stimuli involving *Quantifiers and negation* (e.g., *There are many red blocks and few red blocks*), and far worse in the other categories, suggesting they may only apply relatively simple linguistic heuristics to reason about the physical query.

To better understand the limitations of our model, we manually inspect stimuli in which our model deviates most from human judgments ($n = 10$ with greatest Wass. Distance). We find two suggestive grounds for future work:

People draw exact logical inferences; our model uses sample-based approximation. Our model consistently deviates from human judgments on stimuli that people can reason about exactly, such as those involving equality (eg. *Half of the blocks are yellow, and half are red.*). Humans produce a sharp, exact judgment, which our model approximates with sample-based inference. These cases are one exception in which the fewshot LLM baseline outperforms our model, generalizing the exact human judgements to new stimuli.

People may pragmatically interpret scene descriptions; our model uses literal semantics. Our model may also deviate from human judgments when people apply an intuitive, pragmatic interpretation to the scene descriptions. Our model translations are based on a ground truth, literal semantics. Humans, however, often appear to pragmatically strengthen descriptions based on assumed relevance of all conditions – in many cases, for instances, people overweight the contribution of blocks described to be on the table edges (eg. *There are more red blocks than yellow blocks on the table, and there are more yellow blocks than red blocks on the edges of the table*) relative to our model, suggesting that people assume the edge is mentioned because it impacts the downstream result.

Perhaps surprisingly, we find that **the model rarely makes obvious semantic translation errors.** In the 10 stimuli that we inspect, we find only one, phrase-level translation error: *There are several stacks of red blocks on the table* is translated to `condition(filter(isRed, world.stacks).length > 1)`, when *several* intuitively suggests an upper and lower threshold. While the model produces literal interpretations, as discussed above, we find no other obviously incorrect translations.

Conclusions and future directions. We conclude with several avenues for future work. One clear next step might translate language that specifies background knowledge or poses arbitrary new queries, broadening the integration of language and physical reasoning. Our results also suggest that integrating this approach with *pragmatic* inference, such as in Frank & Goodman (2012), is crucial for capturing a human-like understanding of language. As a cognitive model, we must consider consider how the joint distribution we instantiate in an LLM can be learned from plausible amounts of data; future work should also evaluate against other LLMs, testing what latent physics can be acquired in itself with more or more targeted supervision. Finally, integrating this approach with perception, using *inverse graphics* (Yi et al., 2018) approaches to construct structured scene representations from perceptual inputs, could broaden this approach to bridge between language, our rich internal physical reasoning, and grounding in the external, perceivable world.

References

Baillargeon, R. (2004). Infants' physical world. *Current directions in psychological science*, *13*(3), 89–94.

Barsalou, L. W., Santos, A., Simmons, W. K., & Wilson, C. D. (2008). Language and simulation in conceptual processing. *Symbols, embodiment, and meaning*, 245–283.

Bartsch, R. (1973). The semantics and syntax of number and numbers. In *Syntax and semantics volume 2* (pp. 51–93). Brill.

Barwise, J., & Cooper, R. (1981). Generalized quantifiers and natural language. *Linguistics and Philosophy*, *4*(2), 159–219.

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... others (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

Carey, S. (2009). *The Origin of Concepts*. Oxford University Press.

Catto, E. (2023). *Box2D: A 2D Physics Engine for Games*. http://box2d.org.

Chater, N., Tenenbaum, J., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, *10*(7), 287–291.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. d. O., Kaplan, J., ... others (2021). Evaluating large language models trained on code. *arXiv*.

Cooper, R., Dobnik, S., Lappin, S., & Larsson, S. (2015). Probabilistic type theory and natural language semantics. *Linguistic issues in language technology*, *10*.

Fodor, J. A. (1975). *The language of thought* (Vol. 5). Harvard university press.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Gelman, R., & Gallistel, C. R. (1986). *The child's understanding of number*. Harvard University Press.

Goodman, N. D., & Lassiter, D. (2015). Probabilistic semantics and pragmatics: Uncertainty in language and thought. *The handbook of contemporary semantic theory, 2nd edition. Wiley-Blackwell*.

Goodman, N. D., & Stuhlmüller, A. (2014). *The Design and Implementation of Probabilistic Programming Languages*. http://dippl.org. (Accessed: 2023-1-31)

Harris, Z. S. (1954). Distributional structure. *Word*. Retrieved from http://psycnet.apa.org/psycinfo/1956-02807-001

Heim, I., & Kratzer, A. (1998). *Semantics in generative grammar*. Malden, MA: Wiley-Blackwell.

Hespos, S. J., & Baillargeon, R. (2008). Young infants' actions reveal their developing knowledge of support variables: Converging evidence for violation-of-expectation findings. *Cognition*, *107*(1), 304–316.

Jackendoff, R. S. (1985). *Semantics and cognition* (Vol. 8). MIT press.

Klein, E. (1980). A semantics for positive and comparative adjectives. *Linguistics and philosophy*, *4*, 1–45.

Lakoff, G. (1988). Cognitive semantics. In U. Eco (Ed.), *Meaning and mental representations* (pp. 119–154). Bloomington: Indiana University Press.

Landau, B., & Jackendoff, R. (1993). Whence and whither in spatial language and spatial cognition? *Behavioral and brain sciences*, *16*(2), 255–265.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. University of Chicago press.

Liu, R., Wei, J., Gu, S. S., Wu, T.-Y., Vosoughi, S., Cui, C., ... Dai, A. M. (2022). Mind's eye: Grounded language model reasoning through simulation. *arXiv preprint arXiv:2210.05359*.

Rips, L. J., & Hespos, S. J. (2015). Divisions of the physical world: Concepts of objects and substances. *Psychological bulletin*, *141*(4), 786.

Schuler, K. K. (2005). *Verbnet: A broad-coverage, comprehensive verb lexicon*. University of Pennsylvania.

Spelke, E. S. (1990). Principles of object perception. *Cognitive science*, *14*(1), 29–56.

Spelke, E. S., Gutheil, G., & Van de Walle, G. (1995). The development of object perception. *Visual cognition: An invitation to cognitive science*, *2*, 297–330.

Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive science*, *12*(1), 49–100.

Téglás, E., Vul, E., Girotto, V., Gonzalez, M., Tenenbaum, J. B., & Bonatti, L. L. (2011). Pure reasoning in 12-month-old infants as probabilistic inference. *science*, *332*(6033), 1054–1059.

Ullman, T., Goodman, N., & Tenenbaum, J. (2012). Theory learning as stochastic search in the language of thought. *Cognitive Development*.

Ullman, T. D., Goodman, N. D., & Tenenbaum, J. B. (2010). Theory Acquisition as Stochastic Search. In *Proceedings of thirty second annual meeting of the cognitive science society*.

Ullman, T. D., Spelke, E., Battaglia, P., & Tenenbaum, J. B. (2017). Mind games: Game engines as an architecture for intuitive physics. *Trends in cognitive sciences*, *21*(9), 649–665.

van Eijck, J., & Lappin, S. (2012). Probabilistic semantics for natural language. *Logic and interactive rationality (LIRA)*, *2*, 17–35.

van Tiel, B., Franke, M., & Sauerland, U. (2021). Probabilistic pragmatics explains gradience and focality in natural language quantification. *Proceedings of the National Academy of Sciences*, *118*(9), e2005453118.

Williamson, T. (2002). *Vagueness*. Routledge.

Yi, K., Wu, J., Gan, C., Torralba, A., Kohli, P., & Tenenbaum, J. B. (2018). Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *arXiv preprint arXiv:1810.02338*.