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 make predictions and draw inferences about physical worlds described entirely in language. Understanding this everyday language requires inherently probabilistic reasoning—over possible physical worlds invoked in language and over uncertainty inherent to those physical worlds. In this paper, we propose **PiLoT**, a neurosymbolic generative model that translates language into probabilistic programs grounded in a physics engine. Our model integrates a large language model (LLM) to robustly parse language into program expressions and uses a probabilistic physics engine to support inferences over scenes described in language. We construct a **linguistic reasoning benchmark** based on prior psychophysics experiments that requires reasoning about physical outcomes based on linguistic scene descriptions. We show that PiLoT well predicts human judgments and outperforms LLM baselines.

1 INTRODUCTION

Physical intuitions pervade everyday language. We can describe and imagine a *tall stack of plates*, a *heavy box*, and objects that *move*, *bounce*, or *collide*. We flexibly make predictions (*what happens if one pushes that table stacked with plates?*) and infer underlying properties of the world (*how heavy is that box that no one can lift?*). Understanding this language requires integrating uncertainty in language (e.g., possible heights picked out by *tall*) with uncertainty about the world itself.

How do we relate the meanings of language to what we know about the physical world? A productive line of computational cognitive models, which is based on extensive developmental evidence, has modeled human physical understanding as probabilistic inference over a *mental physics engine*, using representations like those for simulations in video games (Battaglia et al., 2013; Ullman et al., 2017). But how are these capabilities integrated with language, allowing us to imagine and draw inferences over possible physical worlds, described in words? Recent AI advances suggest one route for modeling human language understanding, using *large language models* (LLMs) (Brown et al., 2020; Chowdhery et al., 2022). Close analysis, however, suggests that these models often fall short in capturing how humans *reason* about language when it requires reasoning about a structured world state (e.g., Collins et al., 2022). Increasingly, a parallel line of work suggests instead augmenting these models with external world knowledge and computational capabilities, such as calculators (Cobbe et al., 2021), knowledge bases (Karpas et al., 2022), and even physics simulations applied towards deterministic, textbook-style questions (Liu et al., 2022).

Here, we consider the computational challenge of understanding language that captures our intuitive, probabilistic understanding of the physical world. We seek a modular account that explains how language is integrated with, but distinct from, our general physical reasoning abilities. This work makes three main contributions towards these ends (Fig. 1). First, we propose a **linguistic physical reasoning benchmark** inspired by an existing battery of visual psychophysics tasks (Battaglia et al., 2013), designed to measure commonsense inferences about physical scenes described in everyday language. Next, we propose **PiLoT** (*Physics in a Language of Thought*), a **computational model that translates language into probabilistic programs grounded in a physics engine**, as a framework for modeling human-like physical reasoning over language. This model builds on a classical

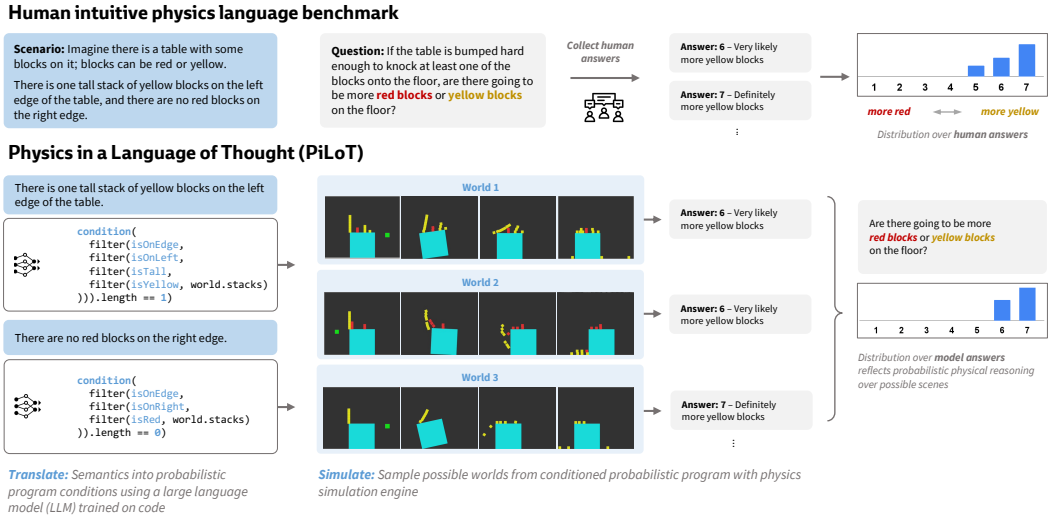

Figure 1: Human language understanding draws on our flexible, intuitive physical knowledge. (*Top*) We collect human judgments about physical outcomes based on descriptions of a tabletop scene with varying configurations of red and yellow blocks.(*Bottom*) Our model, **PiLoT**, reasons about these descriptions by translating language into *probabilistic program expressions* that condition a generative model over possible scenes using a code LLM. To answer questions about physical outcomes, PiLoT samples and simulates scenes from the conditioned model using a *physics engine*, producing inferences that correlate well with human judgments.

theoretical tradition that suggests we construct linguistic meaning from cognitive representations in a compositional *language of thought* (Fodor, 1975; Jackendoff, 1985; Lakoff, 1988), and more recent proposals that address uncertainty in meaning using probabilistic semantic representations (van Eijck & Lappin, 2012; Cooper et al., 2015; Goodman & Lassiter, 2015). One outstanding challenge for scaling these approaches, however, has been implementing broad-coverage functions that can map generally between human language and an underlying semantics. Moreover, prior work has left largely open how the semantics of language can interface formally with physical knowledge. In this paper, we propose a computational framework that addresses the first challenge using *large language-code models* to translate between sentences in language and symbolic semantic expressions, and show that this approach can generalize across a broad range of sentences. Then, by modeling these semantics as *probabilistic programs grounded in a physics engine*, our model can flexibly construct general, structured meanings that also support simulation and physical inferences over language. When applied to our linguistic benchmark, we show that PiLoT **robustly predicts human reasoning about linguistic physical scenes**. Our model better correlates with human judgments and outperforms the directly-queried LLM baseline across the benchmark as a whole. We also find that our model **best predicts the underlying distribution of human judgments**, capturing the uncertainty inherent to how we reason about abstract, linguistic descriptions about these scenes.

2 Linguistic Physical Reasoning Benchmark

We propose a linguistic and physical reasoning task inspired by psychophysics stimuli from 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 benchmark adapts this domain to scenes described in *language*. Unlike visual images, this task requires reasoning over the additional uncertainty inherent to language.

Each stimuli in our benchmark begins with a linguistic description of the general domain of scenes (*Imagine a table with some red or yellow blocks on it*) and 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?*

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. (Montague, 1973; Barwise & Cooper, 1981; van Tiel et al., 2021).
- **Gradable adjectives**: adjectives describing the stacks as *tall*, *very tall*, or *short* (Klein, 1980; Williamson, 2002; Lassiter & Goodman, 2017).

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

Human linguistic reasoning experiment: We 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 recruited 160 human participants from the Prolific platform and collected approximately 40 human responses per stimulus.

3 Our model: PiLoT

In this work, we set out to architect a cognitive model of physical reasoning inspired by theories of mental simulation and the principle of modularity. The resulting model, which we call PiLoT, consists of three modules: A probabilistic generative model over possible scenes, a physics simulator, and a language-to-code translation model. Together, the generative model and physics simulator implement a version of the model used in Battaglia et al. (2013). Meanwhile, the translation model extends their framework to integrate natural language, in the spirit of Goodman & Lassiter (2015).[1]

Probabilistic generative model: We begin by defining a base generative model over possible worlds. We write this model in WebPPL, a JavaScript-based probabilistic programming language (PPL) (Goodman & Stuhlmüller, 2014). Each sample from the model is a stochastically generated initial configuration of blocks. More details about the model can be found in Appendix B.

Physics simulator: To dynamically model scenes sampled from the generative model, we interface the base WebPPL model with a *physics simulator* implemented with 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.[2]

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 base generative model and the query are themselves expressed in WebPPL code, the same methods could be extended 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). Queries to `Codex` (`code-davinci-002`) were issued via the OpenAI API with the temperature parameter set to 0 to ensure that translations adhered to domain semantics and facilitate reproducibility. For each task, we automatically constructed a prompt by concatenating the generative model code and 10 randomly-sampled examples from our domain, each manually annotated with code translations. We found `Codex` a highly

[1] The full code of the model is available at `https://tinyurl.com/phys-lang`.

[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 benchmark yielded marginal improvements of $R^2 < 0.04$ relative to the naive settings that were used in our experiments.

adept translator for our domain, requiring little prompt engineering to produce robust translations of non-trivial phrases.

4 EXPERIMENTS

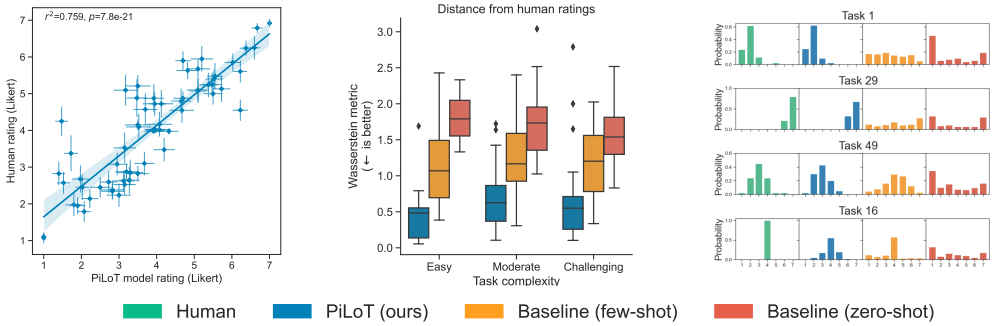

Figure 2: Comparison of PiLoT and baseline models to human ratings at increasing levels of granularity. Left: PiLoT broadly correlates with human Likert ratings across the 64 tasks in our benchmark. Middle: At each task complexity, PiLoT achieves closer fidelity to human ratings than the two baselines, 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 Appendix A for the descriptions associated with each task.)

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

PiLoT: 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 $n = 10$ sampled 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.

Zero-shot LLM: This baseline directly prompts an LLM (Codex) with the exact linguistic setup provided to subjects in the human experiment. 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*) example pairs randomly sampled from heldout stimuli and human responses.

Our model best predicts human judgments across the physical language benchmark. 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 on the benchmark overall (Fig. 2, $R^2 = 0.759$, $p < 0.001$). For baseline models, 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) in Appendix C shows that our model greatly outperforms both baselines. We discuss our experimental findings in more granular detail in Appendix C.

5 CONCLUSION

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* interpretation and inference, such as in Frank & Goodman (2012), is crucial for capturing a human-like understanding of language. Finally, integrating this approach with perception, using *inverse graphics* (Yi et al., 2018) methods to construct structured scene representations from perceptual inputs, could broaden this modeling framework to bridge between language, our rich internal physical reasoning, and grounding in the external, perceivable world.

REFERENCES

Renate Bartsch. The semantics and syntax of number and numbers. In *Syntax and Semantics Volume 2*, pp. 51–93. Brill, 1973.

Jon Barwise and Robin Cooper. Generalized quantifiers and natural language. *Linguistics and Philosophy*, 4(2):159–219, 1981.

Peter W Battaglia, Jessica B Hamrick, and Joshua B Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110(45):18327–18332, 2013.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Susan Carey. *The Origin of Concepts*. Oxford University Press, 2009.

Erin Catto. Box2D: A 2D Physics Engine for Games. http://box2d.org, 2023.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv*, 2021.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training Verifiers to Solve Math Word Problems, November 2021.

Katherine M. Collins, Catherine Wong, Jiahai Feng, Megan Wei, and Joshua B. Tenenbaum. Structured, flexible, and robust: Benchmarking and improving large language models towards more human-like behavior in out-of-distribution reasoning tasks, May 2022.

Robin Cooper, Simon Dobnik, Shalom Lappin, and Staffan Larsson. Probabilistic type theory and natural language semantics. *Linguistic issues in language technology*, 10, 2015.

Jerry A Fodor. *The language of thought*. Harvard university press, 1975.

Michael Frank and Noah D Goodman. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998, 2012.

Rochel Gelman and Charles R Gallistel. *The child's understanding of number*. Harvard University Press, 1986.

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

Noah D Goodman and Andreas Stuhlmüller. The Design and Implementation of Probabilistic Programming Languages. http://dippl.org, 2014.

Ray S Jackendoff. *Semantics and cognition*. MIT press, 1985.

Ehud Karpas, Omri Abend, Yonatan Belinkov, Barak Lenz, Opher Lieber, Nir Ratner, Yoav Shoham, Hofit Bata, Yoav Levine, Kevin Leyton-Brown, Dor Muhlgay, Noam Rozen, Erez Schwartz, Gal Shachaf, Shai Shalev-Shwartz, Amnon Shashua, and Moshe Tenenholtz. MRKL Systems: A modular, neuro-symbolic architecture that combines large language models, external knowledge sources and discrete reasoning, May 2022.

Ewan Klein. A semantics for positive and comparative adjectives. *Linguistics and philosophy*, 4: 1–45, 1980.

George Lakoff. Cognitive semantics. In Umberto Eco (ed.), *Meaning and Mental Representations*, pp. 119–154. Bloomington: Indiana University Press, 1988.

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

Daniel Lassiter and Noah D Goodman. Adjectival vagueness in a bayesian model of interpretation. *Synthese*, 194(10):3801–3836, 2017.

Ruibo Liu, Jason Wei, Shixiang Shane Gu, Te-Yen Wu, Soroush Vosoughi, Claire Cui, Denny Zhou, and Andrew M Dai. Mind's eye: Grounded language model reasoning through simulation. *arXiv preprint arXiv:2210.05359*, 2022.

Richard Montague. The proper treatment of quantification in ordinary english. In *Approaches to natural language: Proceedings of the 1970 Stanford workshop on grammar and semantics*, pp. 221–242. Springer, 1973.

Tomer D Ullman, Elizabeth Spelke, Peter Battaglia, and Joshua B Tenenbaum. Mind games: Game engines as an architecture for intuitive physics. *Trends in cognitive sciences*, 21(9):649–665, 2017.

Jan van Eijck and Shalom Lappin. Probabilistic semantics for natural language. *Logic and interactive rationality (LIRA)*, 2:17–35, 2012.

Bob van Tiel, Michael Franke, and Uli Sauerland. Probabilistic pragmatics explains gradience and focality in natural language quantification. *Proceedings of the National Academy of Sciences*, 118 (9):e2005453118, 2021.

Timothy Williamson. *Vagueness*. Routledge, 2002.

Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Joshua B Tenenbaum. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *arXiv preprint arXiv:1810.02338*, 2018.

A LINGUISTIC PHYSICAL REASONING BENCHMARK EXAMPLES

Easy (1 concept)	*There are four stacks of red blocks, and there is one stack of yellow blocks.* **[Numbers]**	Task 1
	There are short stacks of red blocks, and there are short stacks of yellow blocks. **[Gradable adjectives]**	Task 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]**	Task 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, gradable adjectives]**	Task 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, gradable adjectives]**	Task 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, gradable adjectives]**	Task 64

Table 1: Example stimuli from our linguistic physical reasoning benchmark, 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.

B PROBABILISTIC GENERATIVE MODEL

The code excerpts presented here have been simplified for legibility. As a reminder, our generative model is written in the probabilistic programming language WebPPL (Goodman & Stuhlmüller, 2014). WebPPL extends the deterministic semantics of JavaScript to allow for functions whose behavior is stochastic. For instance, to construct a new block stack, our generative 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 / 2,
            (worldWidth / 2) + table.width / 2)
  )
}
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 >= y_threshold_tall
}
var isOnLeft = function (obj) {
  return obj.x <= x_threshold_left
}
var isNear = function (obj1) {
  return function (obj2) {
    return abs(obj1.x - obj2.x) <= x_threshold_near
  }
}
```

```
// There is a tall stack of red blocks on the left side of the table.
condition(filter(isTall, filter(isRed, filter(isOnLeft, world.stacks))).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 block-worlds 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.

C DETAILED EXPERIMENTAL RESULTS

	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**		**Quantifiers**		**Gradable 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: Benchmark performance of PiLoT and baseline models in comparison to humans, showing Pearson's R^2 and Wasserstein distance (WD) from human ratings. PiLoT outperforms both baselines across the board. 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.

By evaluating our model and baselines in comparison to human performance across the full linguistic reasoning benchmark (Table 2, *Overall*), we find:

Our model best captures the *distribution* of human judgments on each stimulus. We 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.

By considering 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* (eg *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.