

## **UC Merced**

# **Proceedings of the Annual Meeting of the Cognitive Science Society**

### **Title**

Evaluating statistical language models as pragmatic reasoners

### **Permalink**

<https://escholarship.org/uc/item/1024f51v>

### **Journal**

Proceedings of the Annual Meeting of the Cognitive Science Society, 45(45)

### **Authors**

Lipkin, Benjamin  
Wong, Lionel  
Grand, Gabriel  
et al.

### **Publication Date**

2023

Peer reviewed

# Evaluating statistical language models as pragmatic reasoners

Benjamin Lipkin<sup>1</sup> Lionel Wong<sup>1</sup> Gabriel Grand<sup>2</sup> Joshua B. Tenenbaum<sup>1,2</sup>

<sup>1</sup>BCS, MIT <sup>2</sup>CSAIL, MIT

{lipkinb, zyzyyva, gg, jbt}@mit.edu

## Abstract

The relationship between communicated language and intended meaning is often probabilistic and sensitive to context. Numerous strategies attempt to estimate such a mapping, often leveraging recursive Bayesian models of communication. In parallel, large language models (LLMs) have been increasingly applied to semantic parsing applications, tasked with inferring logical representations from natural language. While existing LLM explorations have been largely restricted to literal language use, in this work, we evaluate the capacity of LLMs to infer the meanings of pragmatic utterances. Specifically, we explore the case of threshold estimation on the gradable adjective “*strong*”, contextually conditioned on a strength prior, then extended to composition with qualification, negation, polarity inversion, and class comparison. We find that LLMs can derive context-grounded, human-like distributions over the interpretations of several complex pragmatic utterances, yet struggle composing with negation. These results inform the inferential capacity of statistical language models, and their use in pragmatic and semantic parsing applications. All corresponding code is made publicly available<sup>1</sup>.

**Keywords:** language models; semantic parsing; pragmatics

## Introduction

Natural language understanding unfolds in context and reflects more than literal interpretation. Such a process is posited to be mediated by a series of inferences, which jointly scrutinize mappings between linguistic structure and mental representations in tandem with the plausibility of resulting interpretations. A sentence as simple as “*Mia is tall*” may be broadly meaningful in of itself, but the range of plausible heights a listener will consider shifts with context that “*Mia plays in the WNBA*” or that “*Mia is a three-year old child.*” These contextual inferences are broadly studied as linguistic *pragmatics* (Wittgenstein, 1953; Searle, 1969; Austin, 1975; Levinson, 1983; Grice, 1989; Clark, 1996).

Recently, work on large-scale training of transformer language models has produced engineering artifacts that perform exceedingly well across a range of natural language processing (NLP) benchmarks. While trained explicitly to optimize an objective of next-token prediction, such systems implicitly recapitulate large swaths of the traditional NLP pipeline, from POS tagging and parsing to semantic role labeling and coreference resolution (Tenney, Das, & Pavlick, 2019; Bommasani et al., 2021). Indeed, a growing body of contemporary work utilizes LLMs to synthesize program-like represen-

tations from natural language (NL) inputs for use in downstream applications from action planning to theorem solving (Acquaviva et al., 2021; Gao et al., 2022; Collins, Wong, Feng, Wei, & Tenenbaum, 2022; Mishra et al., 2022; Zelikman, Huang, Poesia, Goodman, & Haber, 2022; Wong et al., prep.). In leveraging such systems as semantic parsers, this work casts LLMs as formal accounts of the mapping between linguistic forms and representations of meaning. However, such evaluations have been largely restricted to *literal* language use and translation. In contrast, *pragmatic* meaning estimation often requires considering a distribution over multiple interpretations in context, presenting additional complexity (Fried, Tomlin, Hu, Patel, & Nematzadeh, 2022; Hu, Floyd, Jouravlev, Fedorenko, & Gibson, 2022; Ruis et al., 2022; Hu, Levy, Degen, & Schuster, 2023).

Existing models of pragmatic reasoning typically rely on explicit probabilistic computation, often within the *Rational Speech Acts* (RSA) communication framework, whereby a pragmatic listener reasons about an informative speaker to infer intended meanings (Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013; Goodman & Frank, 2016). We ask: can statistical language models *amortize* common pragmatic inferences, recovering approximately equivalent distributions between language and contextually-modulated meanings?

To address this question, in this work, we explore the case of interpretation over the gradable adjective “*strong*” in describing a player in a fictional game. Conditioned on context describing a generative model over possible worlds, expressing a numerical prior on “*strength*”, among other variables, our paradigm invokes estimation over numerical interpretations of textual descriptions of a novel player’s strength. We collect both LLM-estimated and human-measured distributions over the interpretations of such utterances, and explore composition with additional dimensions of complexity. We find that LLMs impressively infer context-aware, human-like distributions over complex pragmatic utterances such as “*very strong for a beginner*”. Simultaneously, we observe a failure to compose such inferred meanings with negation, e.g., “*not strong*” or polarity inversion, e.g., “*weak*”, offering insights into potential shortcomings.

## Meaning as probabilistic programs

In expressing formal representations of linguistic meaning, one approach has been to build from the framework of model

<sup>1</sup><https://github.com/benlipkin/probsem/tree/CogSci2023>

theoretic semantics (Kripke, 1963; Montague, 1973; Partee, ter Meulen, & Wall, 1990; Kratzer & Irene, 1998), in combination with uncertainty quantification (Van Eijck & Lappin, 2012; Cooper, Dobnik, Lappin, & Larsson, 2015), converging upon probabilistic programming languages (PPLs), like Church (Goodman, Mansinghka, Roy, Bonawitz, & Tenenbaum, 2012), as a useful substrate. Goodman and Lassiter (2015), in particular, present a framework, which we build from here, for NL as belief updating over probabilistic programs. Starting from a generative model over possible worlds describing a domain, sentences are incrementally expressed as conditioning statements and executed to update posterior beliefs over world states.

Goodman and Lassiter motivate this framework by providing examples through a discussion of a fictional game of tug-of-war (ToW). In this simplified version of the classical children’s game, two teams, each with one or more players, compete against each other, with the winner decided by the team whose players exert the most strength (Goodman & Tenenbaum, 2010; Gerstenberg & Goodman, 2012; Goodman, Tenenbaum, & Gerstenberg, 2014). Starting from this base, Goodman and Lassiter built examples of PPL-mediated contextual semantic analysis. For example, “*Team A has more than 3 players*” could be expressed as `(condition (> (length team-a) 3))`, and when queried if “*Team A*” might beat “*Team B*” (which perhaps has only 2 players), this information would be considered in evaluating the distribution over outcomes of such a match. In elevating this approach beyond literal language use, to scenarios where NL presents with nondeterministic interpretation, Goodman and Lassiter proposed leveraging explicit probabilistic computation via RSA. One difficulty with this framework is the need to manually synthesize programs expressing the semantics of evaluated NL. Drawing from successful approaches in semantic parsing and program synthesis, such a process lends itself increasingly to automation using LLMs.

### Present study

Goodman and Lassiter (2015) have highlighted the elegant capacity of PPLs in expressing the logical representation of sentence meaning, but have left open how such programs be derived in the first place. In parallel, modern semantic parsing work has painted a picture of LLMs as systems capable of mediating such a translation. However, when it comes to scenarios where this task moves beyond literal language use, it is unclear: a) if LLMs are appropriately suited to mediate such sophisticated inferences and b), whether such model estimates would be in line with human expectations. In addressing these questions, we build from the ToW domain model and pursue gradable adjectives as an expressive test bed.

Gradable adjectives, such as “*strong*”, present with vagueness as they lack precise class boundaries. Several approaches have been developed to express the semantics of gradable adjectives, and in one common approach, a free threshold variable is introduced such that “*strong*” be defined as having “*strength*”  $> \theta$  (Cresswell, 1976; Klein, 1980;

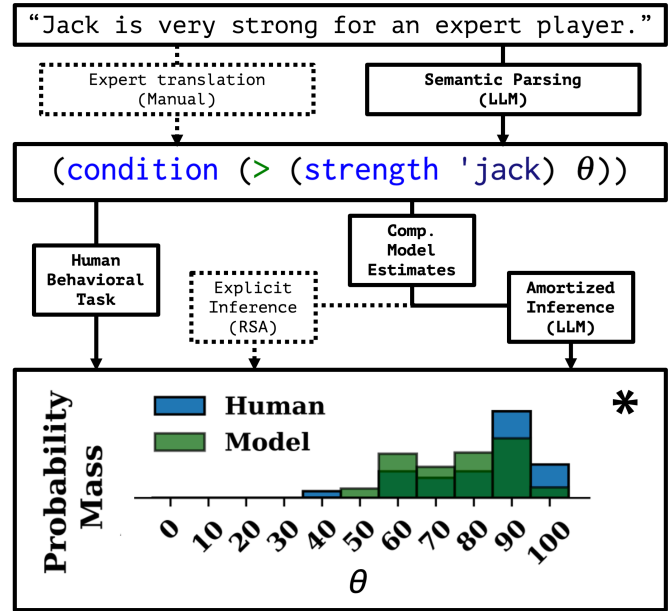


Figure 1: Schematic overview. LLMs stand in for the traditional pragmatics pipeline, often recovering human-like estimates over multiple interpretations of complex constructions.

Kennedy, 2007; Lassiter & Goodman, 2017; Tessler, Tsvilodub, Snedeker, & Levy, 2020). While the distribution over  $\theta$  or other formulations can be derived to various degrees using the recursive probabilistic inference of RSA (Qing & Franke, 2014; Tessler & Goodman, 2022), here we ask whether an LLM can stand in, directly estimating the distribution over  $\theta$  in a single forward pass. See Figure 1 for an overview.

Within the context of ToW players, with a prior over “*strength*” defined in the domain description, we begin with the basic evaluation of “*strong*” and its inverse polarity counterpart “*weak*”, then extending to the inclusion of negation, e.g., “*not strong*”, qualifiers, e.g., “*pretty strong*”, and comparison classes, e.g., “*strong for a novice player*”. In considering the plausibility of LLM inferences, we collect human behavioral norms for the same stimuli to quantify where and how the model captures or fails to reflect human intuitions. We find, on the positive end, that LLMs can perform rather sophisticated contextual amortization of a stack of inferences that include both literal and pragmatic ones, elegantly parsing over complex pragmatic utterances, conditioned on text expressing a generative world model as a probabilistic program. On the negative end, however, LLMs can struggle with the otherwise logically simpler properties of negation or polarity inversion, deviating from human interpretations in such cases. These results inform our understanding of the inferential capacity of LLMs, and as such simultaneously inform debates surrounding the capacities of statistical language learners (see e.g., Piantadosi (2023)).

## Methods

To explore the questions outlined thus far, we begin by more formally defining the ToW domain model in Church, outlining the priors and constraints placed on the semantics explored for the remainder of this work. We define a scoring function, by which the LLM can estimate the probability of particular text interpretations, conditioned on the domain model context and an NL query. To evaluate the efficacy of this scoring function, we developed a set of test materials to evaluate human and LLM-based interpretations of gradable adjectives, and tested our modeling framework and 60 human participants on two variations of this task, one focused primarily on qualification and one on class comparison. Negation and polarity inversion were also explored as part of the qualification experiment. Finally, we consider the distributions over interpretations estimated by the model with respect to those empirically measured in human participants.

### Domain Model and LLM Context

```
;; This Church program models a tug-of-war game between teams of players.
;; Each player has a strength, with strength value 50 being about average.
(define strength (mem (lambda (player) (gaussian 50 20))))
;; Each player has an intrinsic laziness frequency.
(define laziness (mem (lambda (player) (uniform 0 1))))
;; The strength of the team is the sum of the player strengths.
;; When a player is lazy in a match, they pull with half their strength.
(define (team-strength team)
  (sum (map
        (lambda (player)
          (if (flip (laziness player))
              (/ (strength player) 2) (strength player)))
        team)))
;; The winner of the match is the stronger team.
;; Returns true if team-1 won against team-2, else false.
(define (won-against team-1 team-2)
  (> (team-strength team-1) (team-strength team-2)))
;; Now, let us translate some user-defined statements.
;; Each statement begins with either `Condition` or `Query`.
;; `Condition` statements provide facts about the scenario.
;; `Query` statements are questions that evaluate quantities of interest.
;; Condition: Jack is strong.
(condition (> (strength 'jack) 50))
```

Figure 2: Example of full text passed to LLM for a single query. The tug-of-war domain model (blue) and task instructions (green) are consistent across all trials. For each evaluated sentence (yellow), the probability of each program (red) is evaluated to return a score for a given interpretation.

As the context for our gradable adjective experiments, we consider the previously introduced domain of tug-of-war. In Figure 2, we present the precise formulation of this domain as a Church program. Conditioned on this context describing the domain, a description of the task at hand (to translate NL into Church), and a particular NL query, an LLM can then act as a generative model over program expressions, with the capacity both to sample next tokens starting from the prompt, or to assign probabilities to predefined programs under the model.

Critically, we see a prior over  $\text{strength} \sim \mathcal{N}(50, 20)$ . While not all other elements of this domain model are required for our downstream tasks, we include full context so as to evaluate efficacy and robustness within a complete world model.

### Task Description

To condition onto our world model that “*Jack is strong*”, expressed as  $(\text{condition } (> (\text{strength } ' \text{jack}) \theta))$ , what value for  $\theta$  is reasonable? While leveraging RSA is one strategy, it quickly grows intractable to accurately estimate such a range for all gradable adjectives, with the combinatorial space further plagued by the possible composition with additional constraints, e.g., “*somewhat strong*”. So we ask: can an LLM amortize inference of this distribution over  $\theta$  in a way that is pragmatically-sensitive and consistent with human inferences? To evaluate this question, we developed a set of stimuli, each referencing gradable adjectives to describe the strength of a fictional athlete named “*Jack*”. These materials were divided among two experiments.

**E1: Qualifiers** In E1, we first evaluated the probability of sentences about Jack’s strength to be interpreted as programs of the form:  $(\text{condition } (> (\text{strength } ' \text{jack}) \theta))$ , for  $\theta$  from 0 – 100, in intervals of 10. These included both cases where Jack is “*strong*” and where he is “*not weak*”, to various degrees. For each sentence,  $P_{\text{model}}(\theta)$  was estimated by the LLM, and  $P_{\text{human}}(\theta)$  was measured from a collection of human participant point estimates. In addition to these test sentences, a control sentence was included of the form, “*Jack has at least average strength*”, which lacks vagueness and has intention of recovering the majority of probability mass at  $\theta = \mu_{\text{strength}} = 50$ . Then, to test robustness to polarity inversion, we developed a parallel set of materials to evaluate Jack’s weakness, considering instead programs of the form  $(\text{condition } (< (\text{strength } ' \text{jack}) \theta))$ . These materials were directly matched to those in the first part of E1, with only the modification of swapping “*strong*” and “*not weak*” to “*not strong*” and “*weak*”, respectively. The full set of 18 materials can be found in Figure 3.

**E2: Comparison Classes** In E2, we extended this evaluation by introducing comparison classes to conditionally refine interpretation. We modified the definition of  $\text{strength}$  in the LLM prompt, to consider a new variable, the  $\text{league}$  of a player, by injecting the following conditional statement (the full updated prompt can be found in the paper repository):

```
(cond
  ((equal? league 'beginner)
   (gaussian 30 20))
  ((equal? league 'intermediate)
   (gaussian 50 20))
  ((equal? league 'professional)
   (gaussian 70 20))
)
```

Here we test, can an LLM use a verbal descriptor of a player to jointly infer their league membership as well as relative

strength within that league? Drawing from a subset of E1, we developed a new set of materials that incorporate these comparison classes. In particular, we preserved the control form: “*Jack has at least average strength*” and the form which deviated most from the mean in Figure 3: “*Jack is very strong*”. We modified each sentence for each league, along three degrees of abstraction: exact match, synonym, and allusion. For example, for the first league, we assessed Jack’s strength for a “*beginner*”, “*novice*”, and “*someone new to the game.*” The full set of 18 materials can be found in Figure 4.

### Human Participant Evaluation

In order to evaluate  $P_{human}(\theta)$  for each stimulus, two behavioral studies were conducted. 60 participants were recruited from Prolific, 30 for E1 and 30 for E2. Participants provided informed consent and were paid approximately \$15 per hour. The experiment requested that participants move a slider to indicate the threshold ( $\theta$ ) on the strength of a fictional athlete named “*Jack*”, based on independent readings of the stimulus sentences. One participant was removed from E1 for self-reported comprehension difficulties. Analyses include only the remaining participants. The experimental source files, including instructions and stimulus materials, are released with the paper repository.

### LLM Scoring Function

In order to evaluate  $P_{model}(\theta)$  for each stimulus, a scoring function was defined over programs varying  $\theta$ . The OpenAI code-davinci-002 LLM (Chen et al., 2021) is used to parameterize a language model, with the capacity to assign conditional probabilities over any string  $x_i \in \mathcal{X}$ . To interpret the score of each program  $y_i \in \mathcal{Y}$  as a normalized probability with respect to the restricted hypothesis space under consideration, the log-probabilities of the considered programs under the LLM are passed through a softmax function with temperature parameter  $\alpha$ , selected independently for each stimulus sentence using leave-one-out cross-validation (LOOCV) as expanded in the following section.

$$P(y_i) = \frac{\exp(\alpha \log P(x_i))}{\sum_{j=1}^n \exp(\alpha \log P(x_j))} \quad (1)$$

In this case where programs differ only in  $\theta$ ,  $P_{model}(\theta_i)$  is approximated as  $P(y_i)$ . These discrete program probabilities form the basis for subsequent analyses.

### Comparing $P_{human}(\theta)$ and $P_{model}(\theta)$

For each of the 36 stimulus sentences, 29 (E1) or 30 (E2) point estimates on  $\theta$  were measured in human participants. From these point estimates, a discrete empirical distribution over the domain 0 – 100, in intervals of 10, was calculated via normalized counts for each stimulus.

$$P_{human}(\theta_i) = \frac{C(\theta_i)}{\sum_{j=1}^n C(\theta_j)} \quad (2)$$

For the same stimulus sentences, a weight was calculated for each program over the same domain. Such weights were

normalized as in Equation 1 with  $\alpha$  selected for each stimulus by minimizing the sum of the Jensen-Shannon distances (JSD; Equation 4) between  $P_{human}(\theta)$  and  $P_{model}(\theta)$  for the remaining  $N - 1$  stimuli per experiment, using the Nelder-Mead downhill simplex method (Nelder & Mead, 1965).

$$\arg \min_{\alpha} JSD(P_{human}(\theta), P_{model}(\theta)) \quad (3)$$

With  $P_{human}(\theta)$  and  $P_{model}(\theta)$  defined, their similarity was calculated using the Jensen-Shannon distance, a metric distance between two probability distributions  $P$  and  $Q$ , where  $M$  is the point-wise mean between  $P$  and  $Q$ , and  $KL$  is the Kullback-Leibler divergence (Lin, 1991).

$$JSD(P \parallel Q) = \sqrt{\frac{KL(P \parallel M) + KL(Q \parallel M)}{2}} \quad (4)$$

In order to evaluate statistical significance of this similarity metric, a nonparametric permutation test was employed. To generate the null distribution, the values of  $P_{human}(\theta)$  and  $P_{model}(\theta)$  were shuffled over  $\theta$  for  $N = 10,000$  iterations and JSD measured for each variant.  $p$ -values were calculated as the count of null samples less than the true JSD normalized by  $N$ . Raw  $p$ -values were controlled for multiple comparisons using False discovery rate (FDR) correction for the number of tests, within each experiment (Benjamini & Hochberg, 1995).

## Results

In order to evaluate whether LLMs can effectively leverage context to accurately infer distributions over linguistic meaning, several experiments were conducted.

### E1: Qualifiers

For descriptions of Jack’s strength, programs of the form (condition (> (strength 'jack)  $\theta$ )) were evaluated over  $\theta$ .  $P_{model}(\theta)$  is presented in green in Figure 3A.

**LLMs make contextually-aware, pragmatically-sensitive inferences over graded adjectives and qualifiers.** For each variation, a qualitatively smooth and interpretable distribution is reflected over  $\theta$ . For “*Jack is strong*” the majority of probability mass falls  $> \mu_{strength}$ , and when “*Jack is very strong*” it shifts further. For the control “*Jack has at least average strength*”, the mass is correctly placed on  $\theta = \mu_{strength}$ .

**LLMs make mostly human-like inferences, but struggle with negation.** On the same Figure 3A, we see  $P_{human}(\theta)$  presented in blue. Remarkably,  $P_{human}(\theta)$  and  $P_{model}(\theta)$  are generally highly overlapping, even often with complex qualifier composition. In fact, such distributions present with significant similarity for all sentences lacking negation (Figure 3A). However, of the sentences including negation, only half of the interpretations are well-aligned.

**LLMs struggle further with polarity inversion.** To further evaluate the robustness of this framework, a follow-up experiment was conducted, exploring inversion in concept polarity. For a collection of sen-

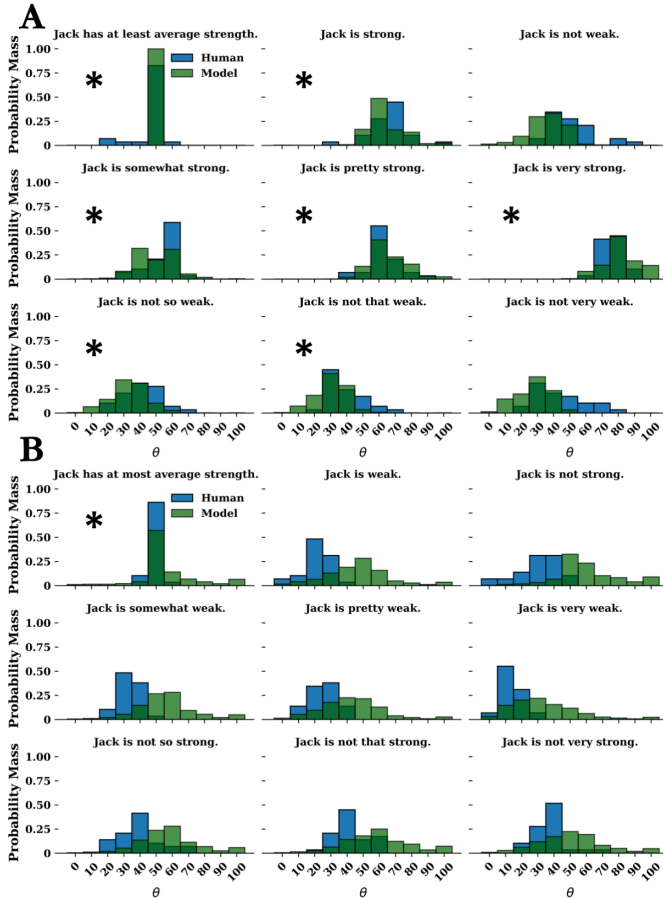


Figure 3: Model-estimated and human-measured distributions over  $P(\theta)$ . Panel A explores programs of the form: (`condition (> (strength 'jack) theta)`), and Panel B: (`condition (< (strength 'jack) theta)`). Each subplot considers a unique sentence, with  $P_{model}(\theta)$  presented in green and  $P_{human}(\theta)$  in blue. An asterisk indicates significant similarity ( $p < 0.05$ ; FDR-corrected) between  $P_{model}(\theta)$  and  $P_{human}(\theta)$ , instantiated as a reduced Jensen-Shannon Distance (JSD; Equation 4) relative to a null permutation analysis.

tences describing the Jack’s weakness, programs of the form, (`condition (< (strength 'jack) theta)`) were evaluated over the domain of  $\theta$ .  $P_{model}(\theta)$  is presented in green in Figure 3B. Once again, distributions appear qualitatively smooth and present with some intuitive characteristics. For example, “*Jack is very weak*” is less than “*Jack is weak*”, and the mean is correctly parsed in the control “*Jack has at most average strength.*” However, a different trend is observed with respect to the alignment with human participants. In this case, where the evaluated concept is of negative polarity with respect to the variable presented in the prompt,  $\theta$  tends to be consistently overestimated by the model. For all sentences other than the control, there is an inability to detect significant similarity between  $P_{model}(\theta)$  and  $P_{human}(\theta)$ .

## E2: Comparison Classes

Selecting the control, “*at least average*”, and the condition deviated most from  $\mu_{strength}$  in Figure 3A, “*very strong*”, a new set of sentences were compiled to describe the strength of “*Jack*” contingent on his membership in different “*leagues*” with individual strength priors. The prompt explicitly presents “*beginner*”, “*intermediate*”, and “*professional*” leagues, with respective means of 30, 50, and 70.

**LLMs accurately parse conditional mixtures, even inferring group membership from indirect descriptors.** Sentences of the form “*Jack ...for a ...*” were presented for each strength description and each league, including both the exact leagues described in the prompt (Figure 4A), as well as previously unseen league descriptors as synonyms (Figure 4B), and even indirect allusions (Figure 4C). Such sentences were parsed and interpreted with outstanding success, significantly aligning with human expectations for 17 of the 18 sentences evaluated, including all control sentences and all sentences at the complexity of direct matches or synonyms.

## Discussion

We began this work with a framework of pragmatic language understanding as an inferential procedure, and next motivated a view of linguistic meaning representation as probabilistic programs. Selecting gradable adjectives as our test bed, we designed a task to evaluate the pragmatic reasoning capacity of LLMs in a complex semantic parsing exercise. Contextualized on code expressing a generative world model defining the semantics of a tug-of-war game, we evaluated a number of sentences about the strength of a fictional player, often composing such sentences with pragmatically complex phenomena. Using an LLM, we estimated  $P_{model}(\theta)$  for each target sentence and conducted human behavioral experiments to empirically measure each corresponding  $P_{human}(\theta)$ .

From our initial evaluation (E1; Figure 3), we learned that LLMs can effectively amortize inference of a smooth distribution over  $\theta$  in a way that is contextually-grounded to the semantics of the prompt and pragmatically-sensitive with respect to gradable adjectives and qualifiers. Such model estimates aligned with human measurements for all descriptions of how “*strong*” a player was, but failed to recapitulate the intricacies of human distributions in the majority of cases where the player was “*weak*”, “*not weak*”, or “*not strong*”. These results suggest that while the model can estimate *some* approximate distribution for each of these cases, the ability to infer an exactly human-like distribution suffers when composing negation in the lexical space, e.g., “*strong*” vs. “*not strong*”, or polarity inversion in the conceptual space, e.g., “*strong*” vs. “*weak*”. This is consistent with prior work noting LLM difficulty in resolving negation more generally (Kassner & Schütze, 2019; Hosseini et al., 2021; Creswell, Shanahan, & Higgins, 2022). It also draws intriguing parallels to child developmental work on concept acquisition, noting observed lags in the mastery of negative polarity concepts, e.g., “*short*”, relative to their positive polarity counter-

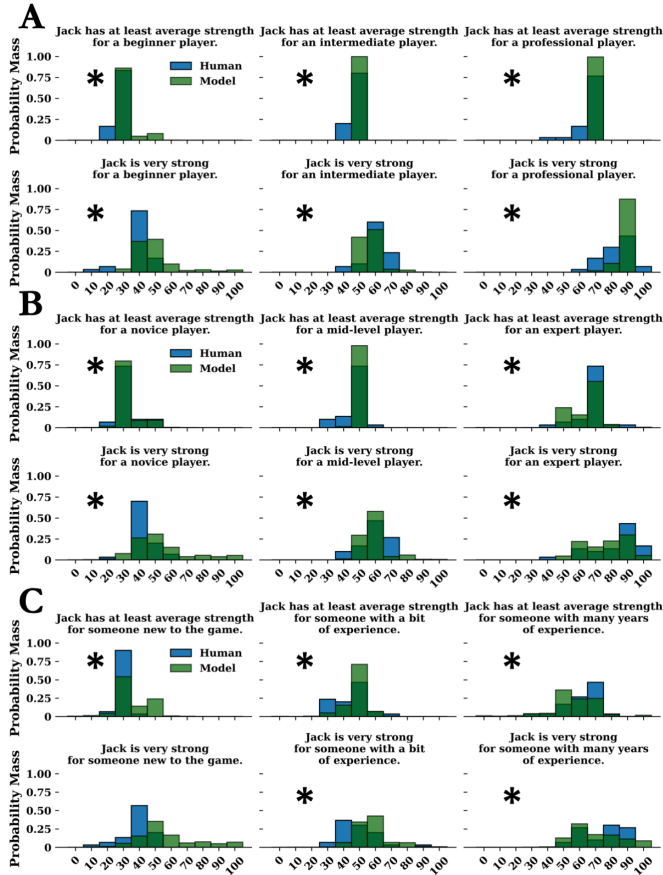


Figure 4: Model-estimated and human-measured distributions over  $P(\theta)$ , incorporating comparison class. Panel A uses exact class from prompt, Panel B: synonyms, and Panel C: allusions. As in Figure 3,  $P_{model}(\theta)$  is presented in green,  $P_{human}(\theta)$  in blue, and an asterisk indicates significant ( $p < 0.05$ ; FDR) similarity between  $P_{model}(\theta)$  and  $P_{human}(\theta)$ .

parts, e.g., “tall”, perhaps highlighting more general asymmetries in concept complexity (Klatzky, Clark, & Macken, 1973; Brewer & Stone, 1975; Barner & Snedeker, 2008). From our evaluation of class comparisons (E2; Figure 4), we further highlighted the context-sensitivity of such models in appropriately resolving conditional mixtures, presenting with impressive robustness in the presence of incorrect references nearby in context. These results are even more powerful when the match between the query and context variable is not exact, but instead needs to be estimated from a synonym or indirect allusion. These results support an argument for the lexical semantic robustness of LLMs under this approach, a convenient case relative to some traditional semantic parsers based on combinatory categorical grammars (CCGs), for which more complex workarounds are often required (Artzi, Das, & Petrov, 2014; Kwiatkowski, Zettlemoyer, Goldwater, & Steedman, 2011; Steedman, 2001).

Overall, these results paint a picture of LLMs as effectively recovering some reasonable distribution in each of these com-

plex test cases, yet highlight some discrepancies with human inferences. If we had perfectly recovered human distributions, this would have led to a series of possible interpretations. One interpretation of such a finding might be that LLMs, just as they appear to implicitly represent other forms of linguistic structure, here implicitly perform inference, as alluded to via other works on amortization (White, Mu, & Goodman, 2020; Wu & Goodman, 2022). Another interpretation could be that, in practice, the statistical regularities of text during training are sufficient to recover these distributions at test time without explicit computation over a world model. Such an account might inform resource-rational frameworks of human language processing, possibly suggesting that partial pragmatic computations could in principle be heuristically approximated, or even retrieved, instead of explicitly recomputed at each instance (Gershman & Goodman, 2014; Gershman, Horvitz, & Tenenbaum, 2015; Dasgupta & Gershman, 2021). While our data do not present LLMs as perfect estimates of human populations across all cases, we believe that these data still at least partially support this second hypothesis. It is indeed possible that some, but not all, of the computations required to solve our task, are amortizable, lending to human-like distributions in some cases, but incorrect approximation in other out-of-domain cases. For example, perhaps composition with negation requires more explicit computation at test time by human participants, which leads to this distributional shift relative to the heuristic estimate of the LLMs. Future work should consider more directly testing this, starting from a framework of computational utility.

**Limitations** While the results presented in this work have proposed a primarily positive image of LLMs as elegantly handling pragmatic inference within a complex semantic parsing task, only a small number of examples within a single scope have been explored thus far. In order to confirm that the conclusions of these results generalize, evaluation of a broader class of pragmatic phenomena in additional task contexts would be required.

**Future Directions** One particularly exciting future direction is connecting LLM-mediated inferences over PPL programs with actual execution of such programs and evaluation of their resulting distributions. If we ask “Can Jill, a very strong beginner, beat Jane, a somewhat strong intermediate?”, such a question can be reduced to neuro-symbolic programming. Leveraging an LLM inference, a distribution over the thresholds on each players’ strengths can be derived. Next, such programs can be explicitly executed in a PPL interpreter, inducing a distribution over each player strength. From this state, it follows easily to query the winner of such a match: (query (won-against '(jill) '(jane))). In then considering more difficult cases, e.g., those involving negation, a hybrid between RSA-like and LLM-mediated approaches might be considered. For example, using LLM estimates to initialize Sequential Monte Carlo (SMC) hypotheses that get updated based on probabilistic program inferences.

## Acknowledgments

We thank our anonymous reviewers for their insightful feedback and recommendations. BL is supported by an MIT Presidential Fellowship and GG by the National Science Foundation Graduate Research Fellowship under Grant No. 2141064. LW and JBT are supported by the MIT Quest for Intelligence, AFOSR Grant #FA9550-19-1-0269, the MIT-IBM Watson AI Lab, ONR Science of AI and DARPA Machine Common Sense.

## References

- Acquaviva, S., Pu, Y., Kryven, M., Sechopoulos, T., Wong, C., Ecanow, G. E., ... Tenenbaum, J. B. (2021). Communicating natural programs to humans and machines. *arXiv preprint arXiv:2106.07824*.
- Artzi, Y., Das, D., & Petrov, S. (2014). Learning compact lexicons for ccg semantic parsing.
- Austin, J. L. (1975). *How to do things with words*. Oxford university press.
- Barner, D., & Snedeker, J. (2008). Compositionality and statistics in adjective acquisition: 4-year-olds interpret tall and short based on the size distributions of novel noun referents. *Child development*, 79(3), 594–608.
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1), 289–300.
- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... others (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.
- Brewer, W. F., & Stone, J. B. (1975). Acquisition of spatial antonym pairs. *Journal of Experimental Child Psychology*, 19(2), 299–307.
- 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*.
- Clark, H. H. (1996). *Using language*. Cambridge university press.
- Collins, K. M., Wong, C., Feng, J., Wei, M., & Tenenbaum, J. B. (2022). Structured, flexible, and robust: benchmarking and improving large language models towards more human-like behavior in out-of-distribution reasoning tasks. *arXiv preprint arXiv:2205.05718*.
- Cooper, R., Dobnik, S., Lappin, S., & Larsson, S. (2015). Probabilistic type theory and natural language semantics. *Linguistic issues in language technology*, 10.
- Cresswell, M. J. (1976). The semantics of degree. In *Montague grammar* (pp. 261–292). Elsevier.
- Creswell, A., Shanahan, M., & Higgins, I. (2022). Selection-inference: Exploiting large language models for interpretable logical reasoning. *arXiv preprint arXiv:2205.09712*.
- Dasgupta, I., & Gershman, S. J. (2021). Memory as a computational resource. *Trends in Cognitive Sciences*, 25(3), 240–251.
- Frank, M., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, 336(6084), 998–998.
- Fried, D., Tomlin, N., Hu, J., Patel, R., & Nematzadeh, A. (2022). Pragmatics in grounded language learning: Phenomena, tasks, and modeling approaches. *arXiv preprint arXiv:2211.08371*.
- Gao, L., Madaan, A., Zhou, S., Alon, U., Liu, P., Yang, Y., ... Neubig, G. (2022). Pal: Program-aided language models. *arXiv*.
- Gershman, S., & Goodman, N. (2014). Amortized inference in probabilistic reasoning. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 36).
- Gershman, S., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, 349(6245), 273–278.
- Gerstenberg, T., & Goodman, N. D. (2012). Ping pong in church: Productive use of concepts in human probabilistic inference. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 34).
- Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, 20(11), 818–829.
- 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., Mansinghka, V., Roy, D., Bonawitz, K., & Tenenbaum, J. B. (2012). Church: a language for generative models. *arXiv*.
- Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science*, 5(1), 173–184.
- Goodman, N. D., & Tenenbaum, J. B. (2010). *Probabilistic Models of Cognition* (First ed.). <http://v1.probmods.org/>.
- Goodman, N. D., Tenenbaum, J. B., & Gerstenberg, T. (2014). Concepts in a probabilistic language of thought. *Center for Brains, Minds and Machines (CBMM) Memos*, 010.
- Grice, P. (1989). *Studies in the way of words*. Harvard University Press.
- Hosseini, A., Reddy, S., Bahdanau, D., Hjelm, R. D., Sordani, A., & Courville, A. (2021). Understanding by understanding not: Modeling negation in language models. *arXiv preprint arXiv:2105.03519*.
- Hu, J., Floyd, S., Jouravlev, O., Fedorenko, E., & Gibson, E. (2022). A fine-grained comparison of pragmatic language understanding in humans and language models. *arXiv preprint arXiv:2212.06801*.
- Hu, J., Levy, R., Degen, J., & Schuster, S. (2023). Expec-



- tations over unspoken alternatives predict pragmatic inferences. *arXiv preprint arXiv:2304.04758*.
- Kassner, N., & Schütze, H. (2019). Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. *arXiv preprint arXiv:1911.03343*.
- Kennedy, C. (2007). Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and philosophy*, 30(1), 1–45.
- Klatzky, R. L., Clark, E. V., & Macken, M. (1973). Asymmetries in the acquisition of polar adjectives: linguistic or conceptual? *Journal of Experimental Child Psychology*, 16(1), 32–46.
- Klein, E. (1980). A semantics for positive and comparative adjectives. *Linguistics and philosophy*, 4(1), 1–45.
- Kratzer, A., & Irene, H. (1998). *Semantics in generative grammar* (Vol. 1185). Blackwell Oxford.
- Kripke, S. A. (1963). Semantical analysis of modal logic in normal modal propositional calculi. *Mathematical Logic Quarterly*, 9(5-6), 67–96.
- Kwiatkowski, T., Zettlemoyer, L., Goldwater, S., & Steedman, M. (2011). Lexical generalization in ccg grammar induction for semantic parsing. In *Proceedings of the 2011 conference on empirical methods in natural language processing* (pp. 1512–1523).
- Lassiter, D., & Goodman, N. D. (2017). Adjectival vagueness in a bayesian model of interpretation. *Synthese*, 194(10), 3801–3836.
- Levinson, S. C. (1983). *Pragmatics*. Cambridge university press.
- Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1), 145–151.
- Mishra, S., Finlayson, M., Lu, P., Tang, L., Welleck, S., Baral, C., ... others (2022). Lila: A unified benchmark for mathematical reasoning. *arXiv preprint arXiv:2210.17517*.
- Montague, R. (1973). 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).
- Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *The computer journal*, 7(4), 308–313.
- Partee, B. B., ter Meulen, A., & Wall, R. (1990). *Mathematical methods in linguistics* (Vol. 30). Springer Science & Business Media.
- Piantadosi, S. T. (2023). Modern language models refute chomsky’s approach to language. *Lingbuzz Preprint, lingbuzz/007180*.
- Qing, C., & Franke, M. (2014). Gradable adjectives, vagueness, and optimal language use: A speaker-oriented model. In *Semantics and linguistic theory* (Vol. 24, pp. 23–41).
- Ruis, L., Khan, A., Biderman, S., Hooker, S., Rocktäschel, T., & Grefenstette, E. (2022). Large language models are not zero-shot communicators. *arXiv preprint arXiv:2210.14986*.
- Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language* (Vol. 626). Cambridge university press.
- Steedman, M. (2001). *The syntactic process*. MIT press.
- Tenney, I., Das, D., & Pavlick, E. (2019). Bert rediscovers the classical nlp pipeline. *arXiv preprint arXiv:1905.05950*.
- Tessler, M. H., & Goodman, N. D. (2022). Warm (for winter): Inferring comparison classes in communication. *Cognitive Science*, 46(3), e13095.
- Tessler, M. H., Tsvilodub, P., Snedeker, J., & Levy, R. P. (2020). Informational goals, sentence structure, and comparison class inference. In *Proceedings of the annual conference of the cognitive science society*.
- Van Eijck, J., & Lappin, S. (2012). Probabilistic semantics for natural language. *Logic and interactive rationality (LIRA)*, 2, 17–35.
- White, J., Mu, J., & Goodman, N. D. (2020). Learning to refer informatively by amortizing pragmatic reasoning. *arXiv preprint arXiv:2006.00418*.
- Wittgenstein, L. (1953). *Philosophical investigations*.
- Wong, L., Grand, G., Lew, A., Andreas, J., Goodman, N. D., Mansinghka, V., & Tenenbaum, J. B. (prep.). Translating from natural language to the language of thought. *preprint*.
- Wu, M., & Goodman, N. (2022). Foundation posteriors for approximate probabilistic inference. *arXiv preprint arXiv:2205.09735*.
- Zelikman, E., Huang, Q., Poesia, G., Goodman, N. D., & Haber, N. (2022). Parsel: A unified natural language framework for algorithmic reasoning. *arXiv preprint arXiv:2212.10561*.