Non-literal Understanding of Number Words by Language Models

Polina Tsvilodub[†*]
Department of Linguistics,
University of Tübingen, Germany

Kanishk Gandhi[†]
Department of Computer Science,
Stanford University

Haoran Zhao[†]
University of Washington

Jan-Philipp Fränken
Department of Psychology,
Stanford University

Michael Franke
Department of Linguistics,
University of Tübingen, Germany

Noah D. Goodman
Departments of Psychology &
Computer Science,
Stanford University

Abstract

Humans naturally interpret numbers non-literally, effortlessly combining context, world knowledge, and speaker intent. We investigate whether large language models (LLMs) interpret numbers similarly, focusing on hyperbole and pragmatic halo effects. Through systematic comparison with human data and computational models of pragmatic reasoning, we find that LLMs diverge from human interpretation in striking ways. By decomposing pragmatic reasoning into testable components, grounded in the Rational Speech Act framework, we pinpoint where LLM processing diverges from human cognition — not in prior knowledge, but in reasoning with it. This insight leads us to develop a targeted solution — chain-of-thought prompting inspired by an RSA model makes LLMs' interpretations more human-like. Our work demonstrates how computational cognitive models can both diagnose AI-human differences and guide development of more human-like language understanding capabilities.

Keywords: hyperbole; pragmatic halo; large language models; pragmatics; Rational Speech Act

Introduction

A friend exclaims, "This coffee cost me a million dollars!" We instantly understand the intended meaning: the coffee was surprisingly expensive (but not a million dollars). Humans often *interpret words non-literally*, effortlessly integrating context, world knowledge, and speaker intent to grasp the meaning behind expressions (Gibbs Jr & Colston, 2006). As large language models (LLMs) become increasingly integrated into our daily lives, three crucial questions emerge: 1) Do LLMs understand literal and non-literal utterances as humans do? 2) Can we use computational models of human cognition to systematically analyze how LLMs interpret non-literal utterances? 3) Can we use cognitive models of human pragmatic language understanding to guide LLMs to interpret meaning in a more human-like way?

In this work, we address these questions by focusing on two common phenomena in the interpretation of *number* words: *hyperbole*, the deliberate use of extreme numerical exaggeration to convey emotion or emphasis (see Ex. (1-a)), and the *pragmatic halo effect*, the tendency to interpret round numbers imprecisely (Ex. (1-b)) and sharp numbers precisely (Ex. (1-c)) (Lasersohn, 1999; Krifka, 2007):

(1) Bob bought a kettle. Bob said:
   a. 'It cost $10000.' ⇝ *Too expensive.* (hyperbole)
   b. 'It cost $50.' ⇝ *It cost around $50.* (imprecise)
   c. 'It cost $48.' ⇝ *It cost exactly $48.* (exact)

Language models trained to auto-regressively predict the next word and subsequently fine-tuned through human feedback have produced impressive performance in many areas (Srivastava et al., 2022, among many others). However, it remains unclear to what extent such training leads to nuanced distinction of literal and non-literal language in LLMs. Recent work has explored non-literal language interpretation in LLMs, from metaphor comprehension (Tong et al., 2021; Liu et al., 2022; Carenini et al., 2023; Prystawski et al., 2023) to pragmatic inference (Jeretic et al., 2020; Jian & Siddharth, 2024; Ruis et al., 2024). While benchmarking efforts have revealed persistent gaps between human and model performance (Sravanthi et al., 2024), we still lack a comprehensive understanding of when and why language models fail at interpreting non-literal language. Understanding these limitations is crucial both for improving models and for insights into how meaning is captured through large-scale language modeling.

To investigate whether LLMs interpret number words in a human-like way, as in Example (1), we compare their interpretations with human judgment data from Kao et al. (2014). Specifically, we elicit LLMs' likelihood estimates for different prices given an uttered number, allowing us to compute the probability of hyperbolic interpretation (*i.e.,* interpreting the price as lower than the stated amount).

The cognitive model of non-literal language interpretation in Kao et al. (2014), suggests that human interpretation is driven by prior knowledge interacting with inferences about the goals of the speaker. We find an interesting disconnect — while LLMs demonstrate human-like prior knowledge about typical prices and what constitutes "expensive", they tend toward more literal interpretations of numerical expressions. This suggests that despite having acquired accurate world knowledge through training, LLMs may lack the pragmatic reasoning mechanisms that humans use to bridge between literal meanings and intended interpretations.

We then explore whether insights from cognitive science toward more human-like interpretations of hyperbole and pragmatic halo, comparing two approaches: cognitive model-inspired chain-of-thought prompting and direct implementation of computational reasoning steps with an LLM. Through

† These authors contributed equally to this work.
*polina.tsvilodub@uni-tuebingen.de

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

Experiment	Prompt
1: Hyperbole & Halo	In each scenario, two friends are talking about the price of an item. Please read the scenarios carefully and provide the probability that the item has the described price. Provide the estimates on a continuous scale between 0 and 1, where 0 stands for "impossible" and 1 stands for "extremely likely". *Daniel* bought a new *electric kettle*. A friend asked him, "Was it expensive?" *Daniel* said, "It cost *$47*." Please provide the probability that the *electric kettle* costs *$50*.
2: Affective Subtext	In each scenario, a person has just bought an item and is talking to a friend about the price. Please read the scenarios carefully and provide the probability that the person thinks that the item is expensive. Provide the estimates on a continuous scale between 0 and 1, where 0 stands for "impossible" and 1 stands for "absolutely certain". *Daniel* bought a new *electric kettle*. It cost *$47*. A friend asked him, "Was it expensive?" *Daniel* said, "It cost *$47*." Please provide the probability that *Daniel* thinks that the *electric kettle* is expensive.
3a: Price Prior	Each scenario is about the price of an item. Please read the scenarios carefully and provide the probability that someone buys the item with the given price. Provide the estimates on a continuous scale between 0 and 1, where 0 stands for "impossible" and 1 stands for "extremely likely". *Daniel* bought a new *electric kettle*. It cost *$50*. Please provide the probability that someone buys the *electric kettle* with this price.
3b: Affect Prior	In each scenario, someone has just bought an item. Please read the scenarios carefully and provide the probability that the buyer thinks that the item is expensive. Provide the estimates on a continuous scale between 0 and 1, where 0 stands for "impossible" and 1 stands for "absolutely certain". *Daniel* bought a new *electric kettle*. It cost *$50*. Please provide the probability that the buyer thinks that the *electric kettle* is expensive.

Table 1: Example prompts used in each experiment. The constant sentences were used as the system prompt. *Italicized* segments varied in each trial.

both methods—providing explicit reasoning chains and implementing Rational Speech Act framework (Goodman & Frank, 2016) computations—we demonstrate that LLMs can achieve more human-like interpretations of non-literal language.

Pragmatic Number Interpretation in Humans

Our work builds on the computational cognitive model developed by Kao et al. (2014) in the Rational Speech Act (RSA) framework, which explains human interpretation of hyperbolic numerical expressions in terms of reasoning about the speaker's communicative intent and prior world knowledge. Specifically, the RSA framework models pragmatic communication as recursive rational reasoning between speakers and listeners (Goodman & Frank, 2016; Degen, 2023). In the basic RSA model, a pragmatic speaker S_1 chooses utterances u to inform a literal listener L_0 of a meaning m, minimizing the listener's surprisal:

$$S_1(u \mid m) = \frac{\exp(\log(P(m \mid [\![u]\!]) - C(u)))}{\sum_{u'} \exp(\log(P(m \mid [\![u']\!]) - C(u')))}$$

where $C(u)$ is the cost of the utterance and $[\![u]\!]$ is the set of meanings compatible with u. A pragmatic listener L_1 then performs Bayesian inference over possible meanings by reasoning about this speaker:

$$L_1(m \mid u) \propto S_1(u \mid m) P(m)$$

where $P(m)$ is the prior probability of a meaning.

To model hyperbolic interpretations like in our coffee example, Kao et al. (2014) extend this framework to capture how a single utterance can convey multiple meanings. Their extended model represents a multi-dimensional meaning space where an utterance about price conveys both the actual price state s (e.g., the literal cost of the coffee) and the speaker's affect a (e.g., that it was surprisingly expensive). The model also incorporates different communicative goals g, allowing the speaker to emphasize either or both of these dimensions:

$$S_1(u \mid s,a,g) \propto \sum_{s',a'} \delta_{g(s,a)=g(s',a')} P(s',a' \mid [\![u]\!]) \cdot e^{-c(u)}$$

The pragmatic listener then interprets the utterance through joint inference over the speaker's goal and intended meaning:

$$L_1(s,a \mid u) \propto \sum_{g} S_1(u \mid s,a,g) P_S(s) P_A(a \mid s) P_G(g)$$

where P_S represents prior beliefs about prices (e.g., how much coffee typically costs), P_A captures the relationship between prices and affect (e.g., when a coffee price would be considered exasperating), and P_G represents the prior over different communicative goals, assumed to be uniform. Kao et al. (2014) showed that this model successfully captures how humans interpret both hyperbolic expressions and the pragmatic differences between round and precise numbers, with model predictions strongly correlating with human judgments.

To explore whether the RSA model can guide LLMs toward more human-like interpretations, we develop a chain-of-thought (CoT) prompt that explicitly walks through key reasoning steps: considering possible speaker intentions, evaluating prior price expectations, and interpreting the utterance accordingly. We demonstrate this reasoning process with an example item (see supplementary) before eliciting the model's interpretation.

Experiments

We closely follow the procedure and the scenarios presented in Kao et al. (2014), about three daily items: an electric kettle, a watch, and a laptop. We study three LLMs in our experiments: GPT-4o-mini, Claude-3.5-sonnet, and and Gemini-1.5-pro. We sample responses from the LLMs with temperature $\tau = 1$ for $n = 10$ times for each query and average predictions across runs.[1]

[1] All materials and data are available at https://sites.google.com/view/pragmatic-lms/home.

	GPT-4o-mini		Claude-3.5 Sonnet		Gemini-1.5-pro	
Prompt	0-shot	1-shot RSA CoT	0-shot	1-shot RSA CoT	0-shot	1-shot RSA CoT
R with humans	0.41	0.579	0.528	0.558	0.365	0.603

Table 2: Correlations between human data and LLM predictions of probabilities of all utterance-meaning ((s,u)) pairs. 0-shot indicates correlations of human results with LLM results under 0-shot prompting, 1-shot RSA CoT indicates correlations of human results with LLM results under one-shot RSA-based CoT prompting.

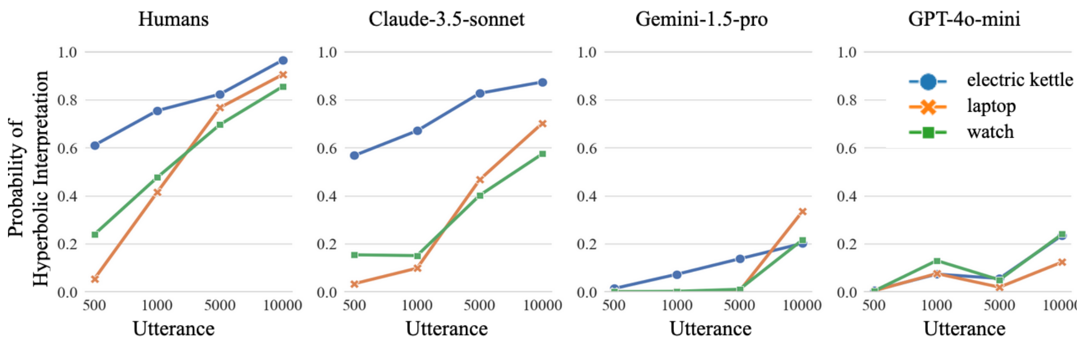

Figure 1: Probability of hyperbolic interpretation, i.e., $u > s$, averaged over sharp and round values of u.

Experiment 1: Hyperbole and Halo

In this experiment, we examine how LLMs interpret price-related utterances, comparing their behavior to human patterns. For hyperbole understanding, we expect LLMs to assign lower probabilities to literal interpretations when they are contextually implausible—for instance, the likelihood of a literal \$10,000 interpretation should be low when discussing an electric kettle's price. To assess pragmatic halo effects, we compare interpretations of sharp versus round numbers, hypothesizing that exact interpretations should be more probable for precise utterances (e.g., "\$51") than for round ones (e.g., "\$50"). To quantify human-likeness, we correlate the LLMs' probability distributions over different price states s given an utterance u with human judgments collected by Kao et al. (2014), for both hyperbolic and halo effects.

Materials and Procedure. The prompts for all experiments were kept as close as possible to original human experiments. Following Kao et al. (2014), we used the following sets of price states and utterances $U = S = \{50 + k, 500 + k, 1000 + k, 5000 + k, 10000 + k\}$, with $k \in \{0,1,2,3\}$ to create exact and round prices. The same procedure was applied to all three items. For each utterance with $u \in U$ of the form "It cost \$$u$.", the LLMs were prompted to predict the probability that the item had each price $s \in S$. The probabilities were then renormalized over S for each $u \in U$.

First, we use a *zero-shot* prompt to generate probabilities of possible price states, given the utterance; examples are presented in Tab. 1 (1: Hyperbole & Halo). Second, we guide the models using a *one-shot chain-of-thought (CoT) prompt* (Nye et al., 2021; Wei et al., 2022). We construct this prompt by translating the computational steps of the RSA model into natural language reasoning for an example scenario (following Prystawski et al., 2023). This RSA-inspired chain-of-thought is appended to our original system prompt as shown in Tab. 1 (1), followed by the context and target task.[2]

Results. Results from our zero-shot evaluation reveal significant disparities between LLM and human interpretations of price-related utterances, as shown in Tab. 2. When examining correlations with human data, we find that LLMs generally default to literal interpretations, with even the best-performing model (Claude-3.5-sonnet) achieving only moderate correlation. Different models exhibited distinct behavioral patterns: GPT-4o-mini tended to assign inflated probabilities to individual utterance-meaning pairs, while Gemini-1.5-pro generally exhibited a bimodal distribution of ratings at the ends of the scale.[3] For hyperbolic interpretation (Fig. 1), we analyzed the probability of hyperbolic meaning by summing probabilities of states where the utterance exceeds the true state ($u > s$), averaging across both round and sharp values (e.g., \$50 and \$49). Only Claude-3.5-sonnet demonstrated a consistent pattern of increasing hyperbolic interpretation probability with higher utterances, though this pattern matched human behavior most closely in the electric kettle domain (cf. Fig. 1, left). Other models consistently underestimated hyperbolic interpretations compared to human benchmarks.

The halo effect analysis (Fig. 2) revealed an even more striking divergence from human behavior. We quantified halo bias by calculating the difference between exact interpretation probabilities ($s = u$) and fuzzy interpretation probabilities ($s \neq u$

[2]The full CoT prompt is in our supplementary materials.
[3]0-shot distributional results are in the supplementary materials.

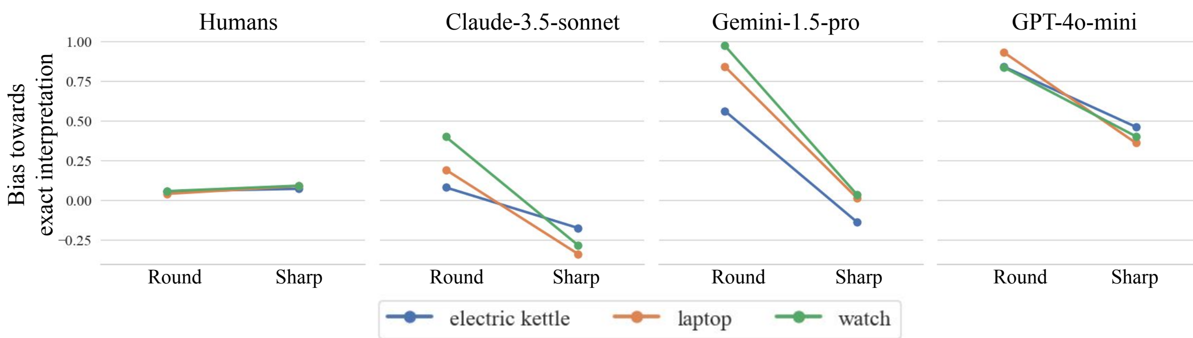

Figure 2: Bias towards pragmatic halo interpretation, calculated by subtracting the probability of a fuzzy interpretation from the probability of the exact interpretation.

and $s \in [u-3, u+3]$). While humans showed a small preference for exact interpretations with sharp numbers, LLMs displayed a large effect in the opposite direction, favoring exact interpretations for round numbers. These findings demonstrate that contemporary LLMs fail to capture human-like pragmatic reasoning in hyperbole and halo interpretation.

RSA-like Chain-of-Thought. We explored if a one-shot Chain of Thought (CoT) prompt, describing the computational process of an RSA model would make LLM responses more human-like. We found that the 1-shot RSA prompt improved correlations between model predictions and human data for GPT-4o-mini and Gemini-1.5-pro (Tab. 2, 1-shot RSA CoT). Interestingly, additional ablation studies showed that prompts explaining only a few components of the RSA model (e.g., mentioning only speaker goals or only price priors) achieved comparable improvements (see supplementary material for details). These ablation results suggest that while RSA-inspired prompting can improve LLM performance, the minimal components sufficient for this improvement differ from the full computational process required to explain human behavior.

Experiment 2: Affective Subtext

Next, we assess the inferred probability that a speaker thinks an item is strikingly expensive (*i.e.,* expressed affect), given the description of the true price s and the speaker's statement u. If the LLMs interpret hyperbole as conveying affect, the likelihood of affect will be higher for hyperbolic utterances (i.e., $u > s$) than for literal utterances (i.e., $u = s$).

Materials and Procedure. We use the same procedure as in Experiment 1 to retrieve the probability of affect, given a *zero-shot prompt* as exemplified in Tab. 1 (2). We use the same sets S and U as in Experiment 1. Following the original experiment, we then round all the states, since we do not predict differences in affect between round and sharp utterances, and calculate the average probabilities of affect for literal utterances (where $s = u$) and hyperbolic utterances (where $u > s$).

Results. Results are shown in Fig. 3. While humans in the experiment from Kao et al. (2014) robustly inferred a distinction between literal and hyperbolic utterances, predicting higher probability of affect given hyperbolic than literal utterances (leftmost facets), LLMs did not. GPT-4o-mini overestimated affect compared to humans, mostly collapsing across literal and hyperbolic utterances. Gemini-1.5-pro treated literal and non-literal utterances more distinctly, but also overestimated affect. Claude was more conservative than humans for hyperbolic utterances, but overestimated affect for literal utterances, with an opposite pattern to humans.

Overall, LLMs did not capture human patterns well when predicting affect probability, given the true price, s, and the uttered price, u. This suggests that LLMs do not map between utterances and affect in a human-like way. This failure may be ancillary to the failure to capture hyperbolic interpretations or may reflect further difficulties with affect (though see Gandhi et al. (2024), where LLMs demonstrate some human-like affective cognition in other contexts).

LLM	GPT	Claude	Gemini
Price prior	0.889	0.93	0.92
Affect prior	0.95	0.973	0.779

Table 3: Correlations of human judgments and different LLM predictions, for price and conditional affect prior probabilities of different prices, across items. These priors were used to fit the respective LM-RSA models in Experiment 3.

Experiment 3: Price and Affect Priors

Given that Experiments 1 and 2 revealed significant differences between LLM and human behavior in processing hyperbole, halo effects, and affective predictions, we designed Experiment 3 to investigate a potential root cause: the accuracy of LLMs' prior knowledge about price distributions and price-affect relationships. Previous work by Kao et al. (2014) has demonstrated that these priors strongly influence human prag-

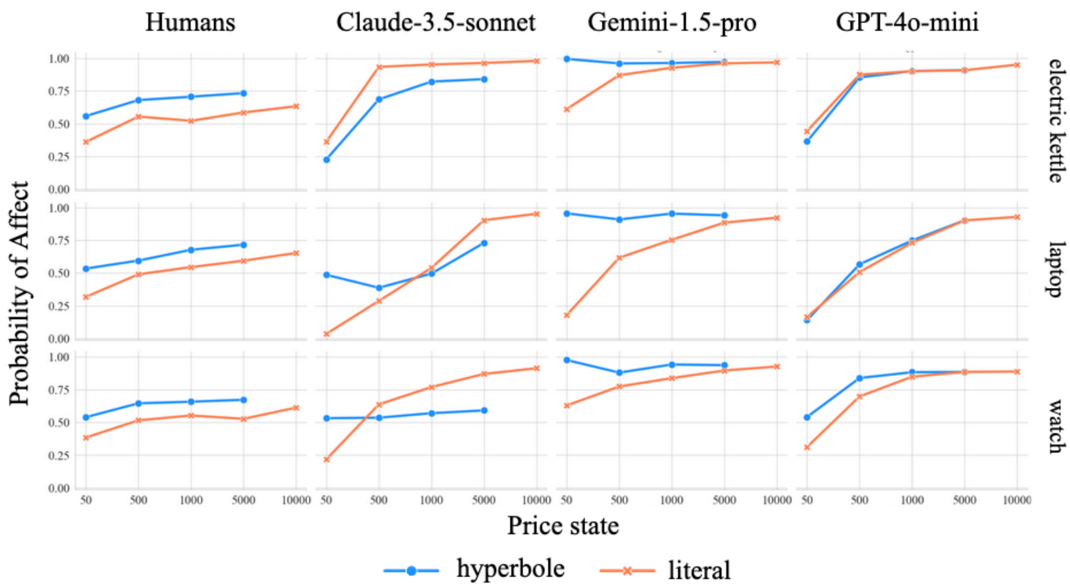

Figure 3: Probability of speaker affect (y-axis), given a price s and an utterance u, predicted by LLMs with zero-shot prompting (with $\tau = 1$) and by humans in Experiment 2 (columns), for different items (rows). Affect is rated by for literal utterances where $u = s$ and hyperbolic utterances where $u > s$.

matic inference. Our experiment focuses on two key aspects: (1) the probability distributions that LLMs assign to different price states for each item, and (2) their predictions of affective responses conditional on item prices. By comparing these model-generated priors against human benchmarks, we can assess whether deficiencies in base knowledge might explain the models' poor performance in pragmatic reasoning tasks.

Strong correlations between LLM and human probability estimates would suggest that pragmatic failures stem from reasoning mechanisms rather than knowledge gaps, while weak correlations would point to fundamental limitations in the models' basic priors about price and affect. By fitting an RSA model using LLM-generated priors, we can quantitatively assess the models' internal consistency—specifically, whether their predictions align with their own stated priors. This analysis provides a systematic approach for isolating reasoning deficiencies independent of the accuracy of the priors themselves. To enable even more precise analysis of the LLMs' priors and reasoning, we additionally fit an RSA model incorporating both LLM-generated priors and conditional utterance likelihoods. If both RSA models demonstrate strong alignment with human data, the reasoning deficiency could be localized specifically to the model's ability to reason about a speaker's intentions.

Materials and Procedure. We use a fixed set of price states $S = \{50 + \text{k},\ 500 + \text{k},\ 1000 + \text{k},\ 5000 + \text{k},\ 10000 + \text{k}\}$, where k was selected from the set $\{0, 1\}$. For price priors, we retrieve LLM predictions with the *zero-shot* prompt asking the LLM to assess the probability of each price $s \in S$ (Tab. 1 (3a)). We renormalize the predictions over all prices S for each item. To retrieve priors over affect, we prompt the LLM *zero-shot* to provide the probability that a person thinks an item is expensive, given the price $s \in S$ of that item (Tab. 1 (3b)). We treat the predictions for each price s as the probability of affect $P(a \mid s)$. To retrieve the conditional probabilities of different utterances, we construct a prompt verbalizing the state and different goals of the RSA speaker S_1.[4] Finally, to investigate to which extent LLMs are consistent with their own priors, we use the predicted *LLM priors* for parameterizing the P_S and P_A in the RSA model. We call the resulting models *LM-RSA* and implement them using a probabilistic programming language WebPPL (Goodman & Stuhlmüller, 2014).

Results. We find that LLM-predicted priors show strong correlation with human data ($r > 0.7$) across price distributions and affect relationships (Tab. 3). This indicates that LLMs possess the prior knowledge that should, in principle, enable them to perform human-like pragmatic inference for both hyperbole and halo effects. We observed a systematic relationship between prior accuracy and zero-shot performance: models with stronger correlations to human priors (progressing from Gemini to Claude) demonstrated correspondingly better zero-shot performance. However, this alignment in prior knowledge, while necessary, proved insufficient to guarantee human-like pragmatic reasoning under prompt-based evaluation.

Do deviations in interpretation nonetheless derive from the small deviations in prior knowledge? We compared models' zero-shot behavior against predictions from RSA models fitted to each LLM's own priors (Fig. 4, upper panel). The results exposed a fundamental inconsistency: LLMs' zero-shot predictions showed relatively weak correlations with their cor-

[4]Details can be found in the supplementary materials.

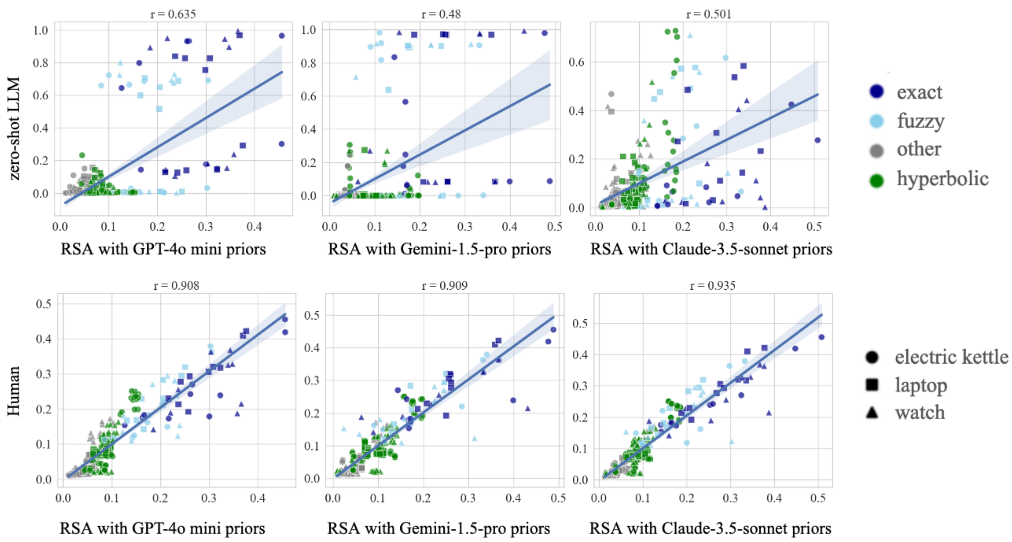

Figure 4: Correlation of predicted probabilities of each pair of (u, s). The plots in the upper panel show predictions of the RSA model with LLM priors (x-axis) against predictions of the same LLM under zero-shot prompting (y-axis). The plot in the lower panel shows predictions of the RSA model with LLM priors (x-axis) against human results (y-axis).

responding RSA predictions, with even the best-performing model (GPT-4o-mini) achieving only moderate correlation (0.635). Yet, these same RSA models showed strong correlations with human judgments (Fig. 4, lower panel) showing that LLM priors are aligned with human priors for price and affect ($R > 0.9$). The LM-RSA models which also included LLM-generated utterance probabilities showed comparable correlations with human judgments ($R > 0.7$).[5] This suggests that the challenge in achieving human-like pragmatic reasoning lies not in the models' priors, but in their ability to systematically apply them during inference.

Discussion

We compared human interpretation data for numerical utterances to LLMs' interpretations, finding substantial differences. This manifests in LLMs' tendency toward literal interpretations, reversed halo effects (preferring exact interpretations for round rather than sharp numbers; Exp. 1), and inconsistent affect attribution between literal and hyperbolic utterances (Exp. 2), despite human-like prior representations (Exp. 3). These findings point to a disconnect in LLM pragmatic reasoning — despite possessing accurate prior knowledge about prices, affect and utterance probabilities — and despite this knowledge being structured in a way that could support human-like inference when processed through an RSA framework — LLMs fail to consistently leverage this information when directly prompted to make pragmatic interpretations.

Our findings highlight an important methodological contribution for understanding LLM behaviors: by systematically decomposing pragmatic reasoning into testable components (priors, affect mappings, utterance likelihoods, and interpretations), we can precisely locate differences between human and AI reasoning. This approach extends beyond traditional behavioral comparisons, allowing us to identify whether differences stem from knowledge gaps or reasoning mechanisms. Such detailed cognitive modeling approaches may prove valuable for understanding other aspects of LLM behavior, particularly in cases where surface-level performance masks deeper processing differences from human cognition. Importantly, our results demonstrate that cognitively-inspired chain-of-thought prompting can help bridge this gap between knowledge and application. We achieved improved correlations with human judgments by decomposing the RSA model's computational steps into natural language reasoning chains. This success suggests that while LLMs may not naturally develop human-like pragmatic reasoning through training alone, they can successfully implement such reasoning when given appropriate computational frameworks that mirror human cognitive processes.

Future research could address important follow-up questions. For instance, frequency effects of different non-literal expressions (cf. McCoy et al., 2024) or potential training modifications to help LLMs better integrate their prior knowledge and context when interpreting hyperbole could be analyzed. Identifying factors that influence how LLMs apply this knowledge in context is also an open question. Our supplementary materials report exploratory analyses that begin to probe these questions through variations in prompting of the models.

Ultimately, our work demonstrates that evaluating LLMs through the lens of cognitive modeling provides a nuanced understanding of how these models deviate from human understanding. By integrating LLMs with cognitive models of pragmatic language use, we can both critically assess the models' internal consistency and provide a framework for improving their performance in interpreting non-literal language.

[5] Full results are provided in the supplement.

Acknowledgments

PT and MF acknowledge support by the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant INST 35/1597-1 FUGG. MF is a member of the Machine Learning Cluster of Excellence at University of Tübingen, EXC number 2064/1 – Project number 39072764. KG was supported by an HAI-SAP Grant and NSF Expeditions Grant Award Number (FAIN) 1918771.

References

Carenini, G., Bodot, L., Bischetti, L., Schaeken, W., & Bambini, V. (2023). Large language models behave (almost) as rational speech actors: Insights from metaphor understanding. In *Neurips 2023 workshop: Information-theoretic principles in cognitive systems.*

Degen, J. (2023). The rational speech act framework. *Annual Review of Linguistics*, *9*(1), 519–540.

Gandhi, K., Lynch, Z., Fränken, J.-P., Patterson, K., Wambu, S., Gerstenberg, T., ... Goodman, N. D. (2024). Human-like affective cognition in foundation models. *arXiv preprint arXiv:2409.11733*.

Gibbs Jr, R. W., & Colston, H. L. (2006). Figurative language. In *Handbook of psycholinguistics* (pp. 835–861). Elsevier.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, *20*(11), 818–829.

Goodman, N. D., & Stuhlmüller, A. (2014). *The Design and Implementation of Probabilistic Programming Languages.* http://dippl.org. (Accessed: 2025-1-25)

Jeretic, P., Warstadt, A., Bhooshan, S., & Williams, A. (2020, July). Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 8690–8705). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.acl-main.768 doi: 10.18653/v1/2020.acl-main.768

Jian, M., & Siddharth, N. (2024). Are llms good pragmatic speakers? *arXiv preprint arXiv:2411.01562*.

Kao, J. T., Wu, J. Y., Bergen, L., & Goodman, N. D. (2014). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, *111*(33), 12002-12007. Retrieved from https://www.pnas.org/doi/abs/10.1073/pnas.1407479111 doi: 10.1073/pnas.1407479111

Krifka, M. (2007). Approximate interpretation of number words: A case for strategic communication. In G. Bouma, I. Krämer, & J. Zwarts (Eds.), *Cognitive foundations of interpretation* (pp. 111–126). Amsterdam: KNAW.

Lasersohn, P. (1999). Pragmatic halos. *Language*, 522–551.

Liu, E., Cui, C., Zheng, K., & Neubig, G. (2022, July). Testing the ability of language models to interpret figurative language. In M. Carpuat, M.-C. de Marneffe, & I. V. Meza Ruiz (Eds.), *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 4437–4452). Seattle, United States: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.naacl-main.330/ doi: 10.18653/v1/2022.naacl-main.330

McCoy, R. T., Yao, S., Friedman, D., Hardy, M. D., & Griffiths, T. L. (2024). Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, *121*(41), e2322420121.

Nye, M., Andreassen, A. J., Gur-Ari, G., Michalewski, H., Austin, J., Bieber, D., ... others (2021). Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*.

Prystawski, B., Thibodeau, P., Potts, C., & Goodman, N. (2023). Psychologically-informed chain-of-thought prompts for metaphor understanding in large language models. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 45).

Ruis, L., Khan, A., Biderman, S., Hooker, S., Rocktäschel, T., & Grefenstette, E. (2024). The goldilocks of pragmatic understanding: Fine-tuning strategy matters for implicature resolution by llms. *Advances in Neural Information Processing Systems*, *36*.

Sravanthi, S., Doshi, M., Tankala, P., Murthy, R., Dabre, R., & Bhattacharyya, P. (2024, August). PUB: A pragmatics understanding benchmark for assessing LLMs' pragmatics capabilities. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Findings of the association for computational linguistics: Acl 2024* (pp. 12075–12097). Bangkok, Thailand: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.findings-acl.719 doi: 10.18653/v1/2024.findings-acl.719

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., ... others (2022). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Tong, X., Shutova, E., & Lewis, M. (2021, June). Recent advances in neural metaphor processing: A linguistic, cognitive and social perspective. In K. Toutanova et al. (Eds.), *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 4673–4686). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.naacl-main.372/ doi: 10.18653/v1/2021.naacl-main.372

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., & Zhou, D. (2022). Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.