On the Same Wavelength? Evaluating Pragmatic Reasoning in Language Models across Broad Concepts

Linlu Qiu*[1] **Cedegao E. Zhang***[1,2]
Joshua B. Tenenbaum[1,2] **Yoon Kim**[1] **Roger P. Levy**[2]
{linluqiu, cedzhang}@mit.edu
[1]MIT CSAIL [2]MIT BCS

Abstract

Language use is shaped by pragmatics—i.e., reasoning about communicative goals and norms in context. As language models (LMs) are increasingly used as conversational agents, it becomes ever more important to understand their pragmatic reasoning abilities. We propose an evaluation framework derived from *Wavelength*, a popular communication game where a speaker and a listener communicate about a broad range of concepts in a granular manner. We study a range of LMs on both language comprehension and language production using direct and Chain-of-Thought (CoT) prompting, and further explore a Rational Speech Act (RSA) approach to incorporating Bayesian pragmatic reasoning into LM inference. We find that state-of-the-art LMs, but not smaller ones, achieve strong performance on language comprehension, obtaining similar-to-human accuracy and exhibiting high correlations with human judgments even without CoT prompting or RSA. On language production, CoT can outperform direct prompting, and using RSA provides significant improvements over both approaches. Our study helps identify the strengths and limitations in LMs' pragmatic reasoning abilities and demonstrates the potential for improving them with RSA, opening up future avenues for understanding conceptual representation, language understanding, and social reasoning in LMs and humans.[1]

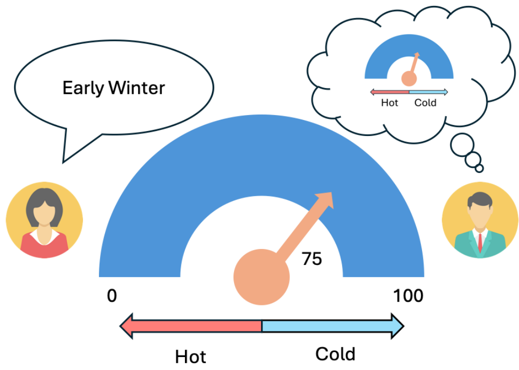

Figure 1: A visual illustration of our experimental setup based on the Wavelength game. In the production task, the speaker (left) is asked to generate a clue ("Early Winter") given the pair of opposite concepts ("Hot" and "Cold") and target value (75). In the comprehension task, the listener (right) is asked to make a guess about the target value given the concepts and clue.

1 Introduction

Human communication occurs *in context*, undergirded by shared goals, norms, and situational cues that shape communication beyond literal meanings of utterances. Within linguistics and cognitive science, *pragmatics* provides a broad framework for studying how speakers and listeners use and interpret language in context (Grice, 1975; Levinson, 1983). Pragmatic communication broadly supposes that given a shared context and communicative goals between speakers and listeners, the speaker chooses utterances to convey specific meanings, while the listener interprets the speaker's intended meanings by assuming that the speaker is being cooperative and informative. This interplay can lead to rich and nuanced phenomena observed in human communication, such as implicature, ambiguity, vagueness, hyperbole, and more (Wittgenstein, 1953; Austin, 1962; Sperber and Wilson, 1986; Clark, 1996).

State-of-the-art language models (LMs) have made significant progress on language understanding and generation, and have now advanced to the point where they are being used actively as conversational agents by millions of people (Wu et al., 2023; Xi et al., 2025). Assessing the pragmatic reasoning abilities of LMs is thus of both theoretical and practical interest. Existing studies of LMs' pragmatic reasoning generally fall into two settings. The first setting adopts a

*Equal contributions.
[1]Code and data are available at https://github.com/linlu-qiu/wavelength-eval.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 19913–19935
November 4-9, 2025

benchmark approach and evaluates LMs on test materials that target various pragmatic phenomena such as presupposition and implicature (Hu et al., 2023; Ruis et al., 2023; Sravanthi et al., 2024). These studies generally find that large LMs can achieve high accuracy and match human error patterns to an extent. The second setting instead asks LMs to act as listeners and/or speakers in communication settings or games and compares the behaviors of LMs against those of human subjects (Jian and Narayanaswamy, 2024; Tsvilodub et al., 2025; Ma et al., 2025b). In this setting, even frontier LMs may not match human performance and demonstrate gaps in their pragmatic reasoning.

We study the pragmatic reasoning capabilities of LMs in the second setting through *Wavelength*,[2] a popular verbal communication game in which one player (listener) tries to guess a hidden number (between 0 and 100) on a scale between two opposite concepts (e.g., "Hot" and "Cold") based on a clue given by another player (speaker), who is trying to communicate the hidden number (see Fig. 1 for an example). While conceptually simple, Wavelength captures several core phenomena crucial for effective pragmatic communication. These include the pursuit of shared goals (measured by the guesser's success), representation of graded meanings (measured on a continuous scale), and application of world knowledge across a diverse set of concepts.

Importantly, Wavelength supports open-ended reasoning about a broad range of words and phrases, but does so in a controlled setting—a property shared by many good games that are useful for studying the mind (Allen et al., 2024). It allows us to collect granular, graded judgments (i.e., numeric values on a continuous scale) from both LMs and humans and ground our quantitative analysis in human distributional data (cf. Ying et al., 2025). In this setting, we study both the comprehension and production sides of pragmatic reasoning, as both are fundamental to conversational agents. We further experiment with incorporating Rational Speech Act (RSA, Goodman and Frank, 2016), a leading computational pragmatics framework, into LMs and assess whether RSA improves model performance or fit to human data.

We test different families of LMs on this dataset and find that model performance systematically increases with model size. On the comprehension task, state-of-the-art models perform at near-human accuracy and show strong correlations with human judgments even with just direct prompting. However, all LMs show significant divergence from humans in terms of the *distributions* of judgments, where we find that human samples are more varied. On the production task, larger models also perform strongly, but more interestingly, we find that RSA-augmented LMs significantly improve upon both direct and Chain-of-Thought (CoT) prompting.

In sum, our main contributions in this paper are:

- A novel dataset collected from human experiments for benchmarking LMs' pragmatic language comprehension and production abilities.
- An LM-based probabilistic inference method that leverages RSA for pragmatic reasoning.
- Evaluations of a range of LMs, finding that larger models perform well in terms of accuracy and human correlations on the comprehension task and LM-RSA reliably helps the production task.

Our study also sheds light on the nature of conceptual representation and language processing in humans and LMs, as we discuss at the end.

2 Background and Related Work

Models of pragmatics. Pragmatics has long been a central subject in the philosophy and science of language (Korta and Perry, 2024). Early theorists have emphasized the social nature of language and analyzed a wide range of interesting or puzzling phenomena in verbal communication that arise because of pragmatics (Searle, 1969; Grice, 1975; Clark, 1996). Formal theories have been developed to target some of them, yet those are not grounded in large-scale empirical studies (Stalnaker, 1978; Heim, 1982; Chierchia et al., 2012; Kamp and Reyle, 2013). More recently, the Rational Speech Act (RSA) represents a promising framework in computational pragmatics that can make quantitative predictions about human pragmatic understanding and reasoning (Frank and Goodman, 2012; Goodman and Frank, 2016; Degen, 2023). At its core, RSA posits a cooperative speaker who selects utterances to maximize conversational utility, and a listener who interprets these utterances through Bayesian inference. It has been used to model a wide range of pragmatic phenomena, including

[2] https://www.wavelength.zone

Left Concept (0)	Target Value	Right Concept (100)	Human-written Clues	Chosen Clue	Human Mean
Deep thought	10	Shallow thought	Evolution, Solving complex problems, Chess, Einstein, Meditation, Quantum mechanics	Solv. complex prob.	11.5
Hot	70	Cold	Coatless, Iced coffee, Refrigerator, Temperature, Colorado mountains, Early winter	Early winter	76.3
Mental activity	100	Physical activity	Running, Sprint, Work out, Race, Gym workout, Marathon	Sprint	93.5
Messy food	20	Clean food	Crawfish, Tacos, Spaghetti, Ribs, Stuffed burrito, Big Mac	Big Mac	25.2
Sport	50	Game	Betting, Darts, Table tennis, FIFA, Mini golf	Table tennis	45.4

Table 1: Example human data. Colored chips represent building blocks of experimental stimuli, and the colors represent the spectrum between 0 and 100. For each production problem we collect human-written clues given the left concept, target value, and right concept. We use human judgments to choose the best clue for the comprehension experiment (chosen clue), and obtain final comprehension judgments (human mean) given the left concept, chosen clue, and right concept. More examples also appear in Table 3; all the concept pairs are presented in Appendix G.

hyperbole, vagueness, generics, and politeness (Kao et al., 2014; Lassiter and Goodman, 2017; Tessler and Goodman, 2019; Yoon et al., 2020).

Pragmatics in language models. There is a growing body of work evaluating the pragmatics in LMs (Zheng et al., 2021; Li et al., 2023; Ruis et al., 2023; Liu et al., 2024b; Sravanthi et al., 2024; Zhao and Hawkins, 2025, see Fried et al. (2023) and Ma et al. (2025a) for surveys). Among them, Hu et al. (2023) is related to our work in that it systematically compares LM and human judgments yet differs in that it utilizes multiple-choice questions rather than graded judgments. Our work shares similarity with Lipkin et al. (2023), which collects graded judgments (on a likert scale) and compares LM and human distributions. But that work only focuses on one domain and studies one LM (OpenAI Codex, Chen et al., 2021). Notably, concurrent works Tsvilodub et al. (2025) and Spinoso-Di Piano et al. (2025) also pursue the direction of incorporating LMs and RSA and ground it in rigorous evaluations, but they focus only on the comprehension side and on a few specific pragmatic phenomena (both on hyperbole and pragmatic halo effects; the latter additionally on irony). On the production side, Jian and Narayanaswamy (2024) evaluates the LM using a reference game and find they are not good pragmatic speakers, consistent with our results. Junker et al. (2025) and Ma et al. (2025b) present benchmarks on multimodal pragmatic expression generation, but they mostly focus on evaluating the vision-language model performance. Murthy et al. (2025) also uses RSA model to study LMs, but focuses on interpret value trade-offs in LMs.

3 The WavelengthEval Dataset

As previously introduced, the Wavelength game captures core features of communication that require pragmatic reasoning. We use it to design a human data collection procedure and create a resulting benchmark for LM pragmatic reasoning evaluation. The core concepts are a pair of opposite or contrasting concepts (spatially *left* and *right*, representing the two extremes and a spectrum in between), a *target value* (between 0 and 100, inclusive), and a *clue* (a single word or a short phrase). We conduct experiments on both the *production* and *comprehension* tasks: for production, the left and right concepts and a target value are given, and the goal is to generate a clue that best communicates the target to another competent language user (maximizing their chance of guessing the value correctly); for comprehension, analogously, the concepts and a clue are given, and the goal is to make the best guess at the intended target value.

These tasks have several properties that make them ideal for human and LM evaluations: (1) the left-right concept pairs are broad and diverse (see examples in Table 1), including both concrete and abstract ones. (2) The choice for clues is open-ended (requiring agents to think over a large set of possible options) but still controlled (only at most a few words are allowed). (3) The target values are precise and graded, making it much harder to blindly guess. (4) The distribution of guesses reflects agents' uncertainty about the problems. These properties match or go beyond well-established paradigms in cognitive science (Griffiths et al., 2024). From a linguistic theory perspective, our test items manifest a range of pragmatic phenomena; examples include scalar adjectives and vagueness ("How cold is cold?"), ad-hoc concepts (using "Einstein" to refer to "deep thought"), and comparison class inference ("Are Jeans considered long—among pants versus all clothing items?") (Kennedy, 2007; Barsalou, 1983; Tessler and Goodman, 2022).

We note that for evaluation purposes, people's comprehension task performance can be measured against the true underlying target value. However, there is no ground truth for judging people's clue generation—the gold standard is to collect human comprehension judgments. In other words, the ground truth evaluation of comprehension is much less costly than that of production. This asymmetry motivates a two-phase experimental design. In Phase 1, we collect a set of human-generated clues for each production problem. Then for each clue we collect a set of human comprehension judgments to filter out the best clue for each problem. The resulting clues form the stimuli for Phase 2, where we expand the set of human judgments for each comprehension problem. The data collection process is as follows, where more details about the human experiments can be found in Appendix A.

Stimuli. We first obtain the official set of pairs of concepts in Wavelength.[3] There are over 200 pairs, and among them we manually choose 50 that are comparatively more commonsensical and less subjective as our stimuli. For each pair, we manually assign two target values that are multiples of 5 between 0 and 100, obtaining 100 problem instances in total. The overall distribution of target values across problems is roughly uniform.

Participants. We recruit 708 human participants in total from the Prolific platform (Palan and Schitter, 2018). The human experiments are approved by the local university IRB. For all experiments we apply the following criteria: adult (default), living in USA, fluent in English, and approval rate 99%+.

Outcome. Our experimental process results in a dataset, WavelengthEval, that contains 50 unique pairs of concepts with 2 target values. For each left-right-value triple, we have an empirically determined high-quality human generated clue and correspondingly 40 human guesses. We then use this dataset to evaluate pragmatic reasoning in LMs through both the comprehension and production tasks, beginning with the former.

[3] https://boardgamegeek.com/thread/2387770/card-list

4 LM Comprehension

Given two opposite concepts and a clue, the language comprehension task requires the listener to guess a value that best represents the clue. An ideal listener should predict a value that is close to the speaker's target. For both humans and LMs, we first estimate their distributions over the scale for each problem and then use the mean as their prediction.

4.1 Methods

We use open-weights, "instruct" versions of LMs for most of our experiments as they provide models of various sizes and access to logits. We consider Llama3 (3.2 3B, 3.1 8B, 3.3 70B) (Dubey et al., 2024), Gemma3 (4B, 12B, 27B) (Gemma Team et al., 2025), Qwen3 (4B, 8B, 32B) (Yang et al., 2025), and DeepSeek-V3 (Liu et al., 2024a) families of models. We also evaluate three representative API-only models: Gemini 2.0 Flash (Gemini Team et al., 2023), Claude 3.7 Sonnet (Anthropic, 2025), and GPT-4.1 (OpenAI, 2025).

We evaluate the LM using three different methods. The first one is *direct* prompting, where we obtain the LM distribution using a prompt that asks it to output responses without intermediate tokens (see Appendix F for prompts). For open-weights models, we approximate this distribution by scoring the next-token probability of all possible values along the scale in increments of 5. For API-only models, we use 32 samples from the LM to estimate this distribution. The second method is *zero-shot CoT* prompting (Kojima et al., 2022), where we instruct the model to think step-by-step before generating the final answer. We similarly use 32 samples to approximate this distribution.

The third method, *LM-RSA*, incorporates probabilistic inference into the LM based on the RSA framework, which views communication as recursive reasoning between a speaker and a listener. We start from the literal listener L_0 (originally representing semantic understanding without pragmatic strengthening), which here is implemented by an LM using either direct or CoT prompting (*Direct-RSA* or *CoT-RSA*). This is based on the assumption that the LM has already acquired the basic meaning of an utterance.

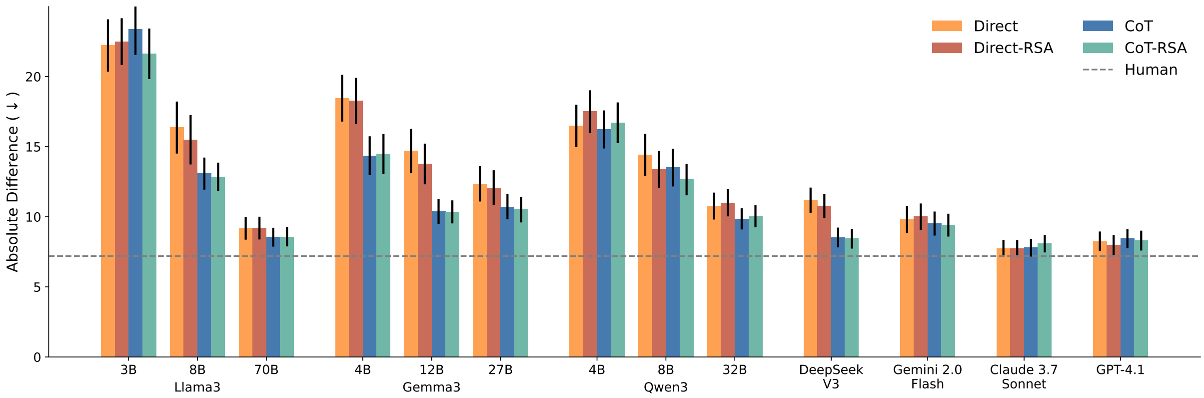

Figure 2: Absolute difference between the model's prediction and the target value. Error bars show standard error over each problem. The dashed line indicates human performance.

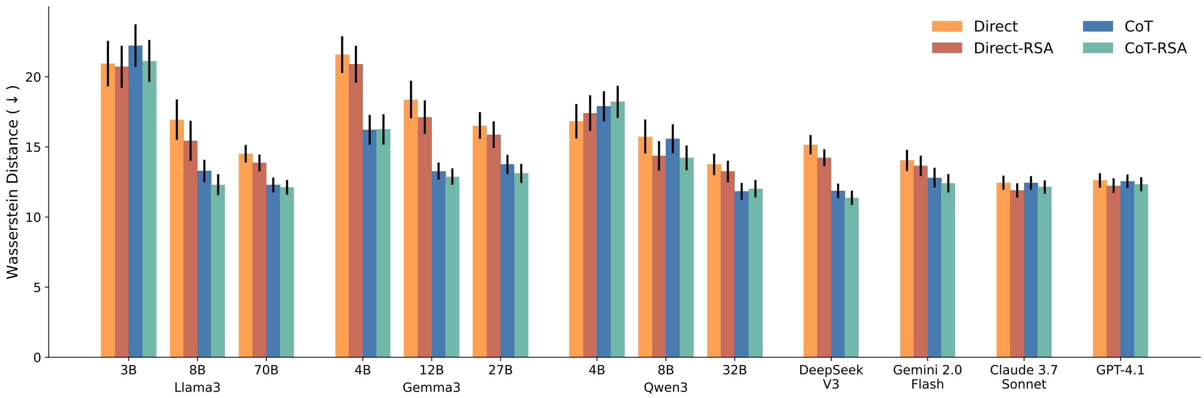

Figure 3: The Wasserstein distance between the model distribution and human distribution. Error bars show standard error over each problem.

In other words, we define

$$P_{L_0}(s \mid u) \propto [\![u]\!](s) \cdot P(s) \approx P_{LM}(s \mid u) \cdot P(s),$$

where s (state) refers to possible target values, u (utterance) refers to possible clues, $[\![\cdot]\!]$ refers to a semantic interpretation function (Heim and Kratzer, 1998), and $P(\cdot)$ refers to the prior over states. Assuming a uniform distribution over all states, equal costs for all utterances, and the softmax rationality parameter $\alpha = 1$, the pragmatic speaker S_1 selects an utterance u given a state s based on:

$$\begin{aligned} P_{S_1}(u \mid s) &\propto \exp(\alpha \cdot U_{S_1}(u; s)) \\ &= \frac{P_{L_0}(s \mid u)}{\sum_{u' \in \mathcal{U}} P_{L_0}(s \mid u')}, \end{aligned}$$

where $\mathcal{U}$ denotes a set of possible utterances, U_{S_1} denotes a utility function, which we define as $\log P_{L_0}(s \mid u)$. Finally, the pragmatic listener L_1 computes the state probabilities given the utterance by performing Bayesian inference over S_1,

$$P_{L_1}(s \mid u) = \frac{P_{S_1}(u \mid s) \cdot P(s)}{\sum_{s' \in \mathcal{S}} P_{S_1}(u \mid s') \cdot P(s')},$$

where $\mathcal{S}$ denotes the set of possible states. This process requires defining the set of alternative utterances, which we use the LM to generate. Specifically, for each state s, we directly sample 1 alternative utterance from the LM without CoT (see prompts in Appendix F).[4] Using the LM as an alternative generator allows us to collect a flexible set of utterances, a task that was challenging in the traditional RSA framework, particularly in our open-ended setting. In Appendix C, we explore a variant of the RSA listener model, which performs similarly to this version.

4.2 Evaluation Metrics

We evaluate the LM from two perspectives. The first is task performance, which evaluates how the LM performs on the comprehension task that requires pragmatic reasoning. We measure this using the absolute difference between the model's prediction and the speaker's target value. Secondly, the comprehension task inherently involves uncertainty. Although individuals may differ on

[4]We explored sampling more alternatives but did not observe significant improvements.

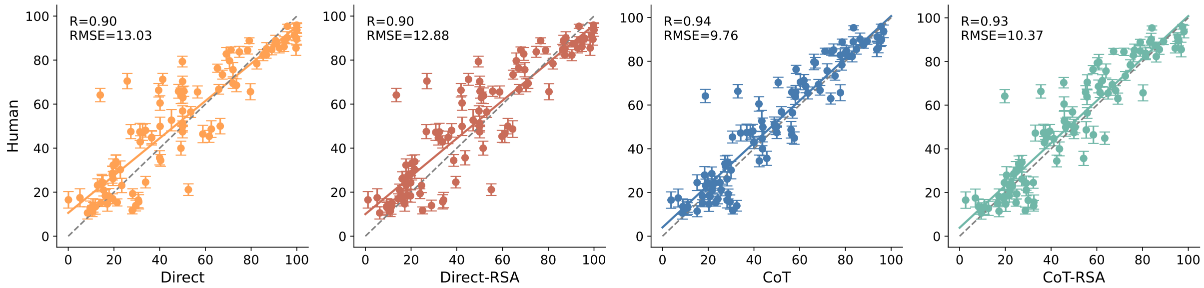

Figure 4: Correlations between model predictions and mean human judgments using Qwen3 32B (see Appendix D.5 for more results). We show Pearson correlations and root mean square standard error (RMSE). Error bars show standard error over 40 human participants.

exact values, there is generally consensus on an approximate range. For example, values in the 60-80 range might all be considered acceptable for "early winter" on the "Hot-Cold" scale. To further evaluate whether the LM captures this uncertainty and maintains a distribution similar to human judgments, we compare the model's distribution with the human distribution using the Wasserstein distance. We also use the Pearson correlations between human and model mean judgments to measure their similarities.

4.3 Results

Task performance. We show the absolute difference between the LM's prediction and target value in Fig. 2. Within each model family, the task performance improves as the model size increases, with the largest model achieving close to human performance. Across model families, the stronger API-only models generally achieve better performance than weaker models. Additionally, CoT improves performance over the direct approach, but RSA does not bring significant improvement.

Humanlike-ness. We show the Wasserstein distance between the model distribution and human distribution in Fig. 3. Similar to the task performance, we find that CoT generally reduces the Wasserstein distance compared to the direct approach. Incorporating RSA helps when using the direct approach, but does not add further value to the CoT prompting. We calculate Pearson correlations between human judgments and the model predictions, as well as the root mean square standard error (RMSE). While smaller models generally show weaker correlation with human judgments, the best model from each model family is typically highly correlated with human judgments, achieving a correlation higher than 0.9 (see Appendix D.5 for details). Fig. 4 shows a breakdown of correlations between the model predictions and human judgments using Qwen3 32B, a representative open-weights model that achieves strong performance. We observe high correlations across different methods, demonstrating that the LM is able to capture mean human judgments to a large extent in this task.

4.4 Analysis

Do LMs and humans perform consistently across different target values? Our target values are roughly uniform from 0 to 100. However, performance on different values may be different for LMs or humans, as certain values might be easier to guess. We show the task performance breakdown by the target value in Fig. 11 (Appendix D). We find that humans perform worse towards the left extreme scale (0), but are generally consistent across different target values. The performance of LMs, however, shows larger variance across target values, especially for small models (e.g., Qwen3 8B, which performs significantly poorly near 0).

LM vs. human distribution. We show examples of the LM and the human distribution in Fig. 5. The top one illustrates the case where using RSA reduces the absolute difference between the model's mean and the target, while also capturing the underlying uncertainty of the human distribution. Intuitively the model becomes less confident on an incorrect value. The bottom one demonstrates a failure case where using RSA hurts performance—the model becomes less confident at extreme values, plausibly because of the other alternatives it generates also assign extreme values.

Choices of alternatives. One hypothesis for why RSA does not improve performance is that, as a listener, the LM tends to be over-confident, producing distributions that are generally concentrated and

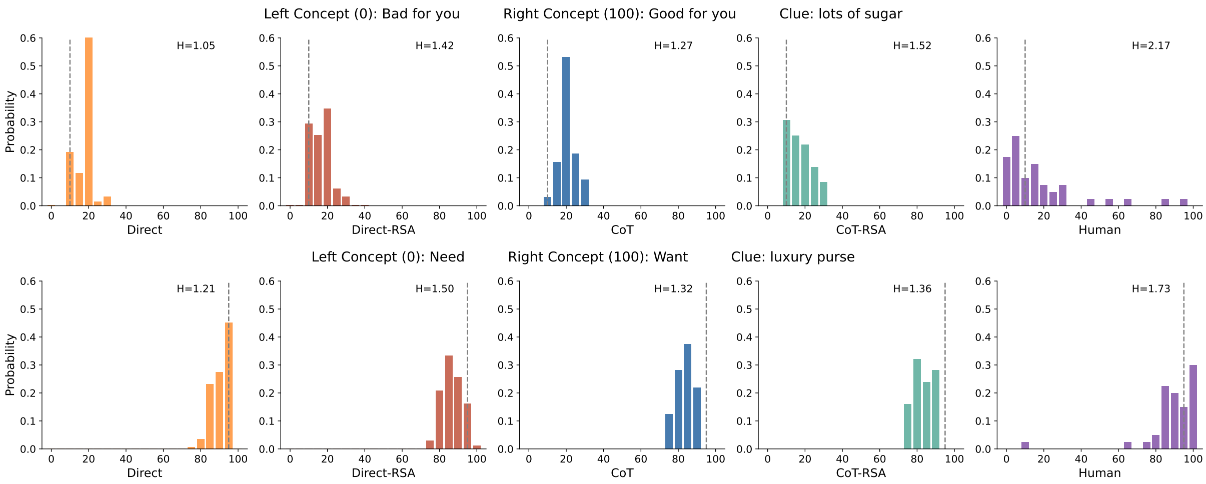

Figure 5: Examples of the LM distribution and human distribution using Qwen3 32B. We show an example where RSA improves the base prompting methods on both task performance and humanlike-ness (top) and an example where RSA hurts (bottom). Dashed lines indicate target values. H refers to the entropy of the distribution.

spiky. In such cases, RSA has limited effectiveness in reshaping the distribution. An alternative hypothesis is that, as a simulated speaker, the LM does not generate good alternatives. We investigate which factor contributes more to the performance by varying the choices of alternatives. We observe that the LM performance is insensitive to the choices of alternatives—providing better or worse alternatives leads to similar performance (Appendix D.3). Therefore, we hypothesize that RSA fails to further improve LM performance on the comprehension task because their concentrated and spiky listener distributions (see Fig. 5 for examples and Appendix Fig. 10 for comparisons of entropy between the LM and human distributions), which limit the benefits of Bayesian re-normalization, the core mechanism of the RSA framework.

Qualitative example. The other potential explanation for why we do not observe significant improvement when using RSA is that the LMs already perform pragmatic inference, either implicitly or explicitly with CoT. We qualitatively examine the reasoning chains of the models and find that in many cases, they already perform RSA-style inference by reasoning about the alternatives, which partially explains the improved performance of CoT compared to the direct approach. We show an example output of Qwen3 32B model in Appendix E.[5]

[5]This only applies to CoT prompting, where we have access to the model's reasoning chains. It is still possible that the model performs similar inference implicitly in the direct prompting case. We leave probing the model's internal mechanisms for future work.

5 LM Production

The language production task requires the speaker to provide a clue that best represents the target value's position on the scale. An ideal speaker, therefore, should provide an informative clue that helps the listener guess the target correctly.

5.1 Methods

We evaluate each LM using three methods. The *direct* and *zero-shot CoT* methods prompt the LM to generate the clue. The former does not explicitly encourage step-by-step reasoning, while the latter does (see prompts in Appendix F).

The *LM-RSA* method instead starts with the literal speaker S_0. We define $P_{S_0}(u \mid s)$ as the LM's distribution $P_{LM}(u \mid s)$; this distribution uses either direct or CoT prompting. Then, the pragmatic speaker S_1 chooses its utterances by considering how listener might interpret their meaning, i.e. sampling the utterance from

$$P_{S_1}(u \mid s) = \frac{P_{S_0}(u \mid s) \cdot P_{L_0}(s \mid u)}{\sum_{u' \in \mathcal{U}} P_{S_0}(u' \mid s) \cdot P_{L_0}(s \mid u')},$$

where P_{L_0} is a direct-prompted LM. This formulation can be viewed as a version of "Inverse-RSA" (Franke, 2022) that starts with a pre-trained speaker and reweights candidate utterances based on pre-trained listener interpretation probabilities (cf. Andreas and Klein, 2016; Hendricks et al., 2016). We use 32 alternative utterances in our experiments.

5.2 Evaluation Metrics

Since the goal of the speaker is to provide clues that helps the listener to guess the target, we eval-

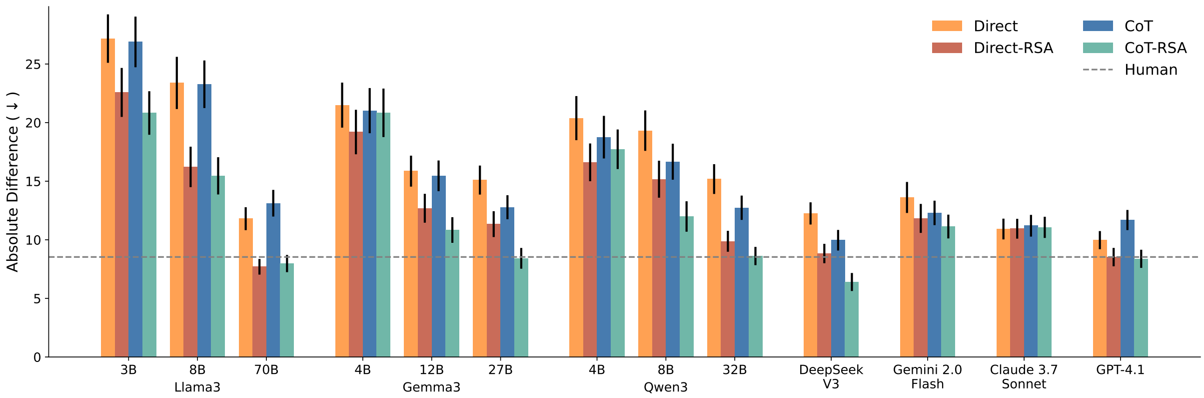

Figure 6: Absolute difference between Qwen3 32B's judgment and the target value, using clues provided by different models. Error bars show standard error over each problem. The dashed line indicates performance using human-written clues.

	Speaker				
Listener (Judge)	Direct	Direct-RSA	CoT	CoT-RSA	Human
Qwen3 32B	14.39	**7.66**	12.25	**6.79**	9.85
Llama3 70B	15.18	**9.87**	12.74	**8.62**	8.54
Human	15.48	**11.92**	14.23	**11.61**	8.98

Table 2: Absolute difference between the prediction and the target value. We compare the performance of Qwen3 32B, Llama3 70B, and humans as listeners (judges), using clues provided by the Qwen3 32B with different methods and human speakers. The best performing methods among the four are bolded (up to statistical significance, $p < 0.05$).

uate the speaker using the listener's performance, i.e., the absolute difference between the listener's prediction and the target value. We use Qwen3 32B with CoT as the listener (i.e., the judge) for most experiments (Zheng et al., 2023), as our results in Section 4 show that it achieves strong performance and exhibits high correlations with humans.

5.3 Results

We show the performance evaluated using Qwen3 32B as the listener in Fig. 6. Similar to findings in the language comprehension task, while smaller models perform poorly in the speaker role, larger models generally achieve strong performance. However, CoT prompting does not necessarily improve performance. By contrast, using RSA consistently improves performance over both direct and CoT prompting, bringing many LMs closer to or even better than human performance, as evaluated by the LM judge.

Note that the performance of Qwen3 32B may not be directly comparable with other models due to the potential issue of self-preference bias where the model favors its own responses (Panickssery et al., 2024; Wataoka et al., 2024). To eliminate this issue, we perform human evaluation on the model's produced clues by collecting 5 human listener guesses for each problem and compute their average. We also use another strong model, Llama3 70B to evaluate its performance. We show results in Table 2. We observe similar trends using human listener and the LM listeners: RSA consistently improves over both the direct and the CoT approach, demonstrating the effectiveness of RSA in improving pragmatic language production. At the same time, on average our human-generated clues are still better than the LM-generated clues when humans are the listener (judge).

5.4 Analysis

Number of alternatives. One important component of RSA is reasoning about how a listener will interpret the meaning of alternative utterances. To investigate the impacts of the *number* of alternative utterances, we evaluate the speaker models using different numbers of utterances. We find that increasing the number of alternatives consistently improves performance (see Appendix D.4).

Qualitative Examples. We show qualitative examples of the LM-generated clues (from Qwen3 32B) and human-written clues in Table 3. In the first example, LM-RSA methods generate intuitively and quantitatively good clues, even arguably better than human's. In the second example, both RSA methods generate the same clue that is clearly better than those of the baselines but not as good as human's. In the third example, CoT-RSA decreases performance, although all the LM-generated clues seem at least decently good. We also show example output of the Qwen3 32B

Left Concept	Right Concept	Target	Direct	Direct-RSA	CoT	CoT-RSA	Human
Hard to remember	Easy to remember	70	Mnemonic (60.4)	Password tip (69.8)	Catchy song (88.4)	Song lyric (62.2)	Mindful (81.4)
Short	Long	60	Jeans (84.4)	Novella (49.0)	Poem (19.6)	Novella (49.0)	Hollywood movie (59)
Art	Commerce	65	Advertising (64.6)	Marketing design (59.4)	Advertising (64.6)	Branding (74.5)	Patreon (62.2)

Table 3: Qualitative examples of the LM-generated clues and human-written clues. The number in brackets shows the mean predictions from 5 human participants.

speaker using CoT prompting in Appendix E. We find that it often reasons about a few alternatives before deciding the final ones, a process similar to what RSA framework aims to model.

6 Discussion

We study the pragmatic reasoning capabilities of LMs across broad concepts through the Wavenlength game. Our results show that simply using prompting, LMs achieve strong performance on both language comprehension and production tasks, and their predictions correlate highly with human judgments. Both task performance and humanlikeness improve as model size increases, suggesting that larger models acquire substantial pragmatic reasoning abilities and rich conceptual knowledge—a potential benefit of scaling. Nonetheless, they also tend to diverge from human judgment distributions: whereas human judgments reflect considerable uncertainty, the LM distributions are generally more concentrated and spiky. As LMs are increasingly used as general assistants, their responses reflecting the uncertainty of the world and capturing the underlying human distribution becomes more important. Future work can investigate methods to improve humanlike reasoning capabilities—including reasoning about human beliefs and intentions—of LMs (Lake et al., 2017; Collins et al., 2024).

Motivated by this consideration, we integrate the classical RSA framework with LMs and demonstrate its effectiveness on the language production task. This integration mitigates certain fundamental challenges inherent to the traditional RSA framework—such as defining the set of alternative utterances and specifying base-level literal listener and speaker models—which have limited its application in more realistic scenarios. It incorporates explicit probabilistic inference into LMs and can be viewed as an instance of neuro-symbolic methods (cf. Wong et al., 2023; Puri et al., 2025). Our approach can improve performance not only with the direct generation setting but also when combined with CoT prompting.

Our WavelengthEval dataset offers data and human judgments to study conceptual knowledge in both humans and LMs. Future work can further extend the dataset by including more target values for each pair of concepts and investigate the conceptual representations along the scale in both humans and LMs (van Tiel et al., 2021; Grand et al., 2022; Tessler and Goodman, 2022). Such extended data with more human-written clues could also provide insights for studying human language processing. For example, when considering the “Cheap” and “Expensive” concepts, some people may reason based on their relative comparison class rather than absolute values, possibly treating “iPhone” as more expensive than “Toyota”, even if their absolute costs might be the opposite (Kamp, 1975; Cresswell, 1976; Kennedy, 2007; Tessler et al., 2020).

More philosophically, our work sheds light on the relationship between acquiring pragmatic knowledge and performing online pragmatic inference. One extreme view of pragmatics is that all of it can be learned, perhaps in a general way (e.g. through next-token prediction) such that there is nothing special about pragmatic phenomena for language acquisition. An opposite extreme view is that pragmatic interpretation and utterance are done entirely on-the-fly based on reasoning about literal meanings given context demands, and literal meanings constitute primary semantic representations in the mind. Our results suggest that there is a plausible middle ground: much of pragmatics can be learned (where literal LM listeners and speakers already perform well), but in many cases explicit reasoning is still desirable (where RSA helps). We leave more systematic development of this position for future work.

7 Conclusion

To sum up, our findings contribute to identifying the strengths and limitations of LM’s abilities in pragmatic language comprehension and production, demonstrate the potential to improve them with RSA, and open up future opportunities to study conceptual knowledge, language use, and social reasoning in LMs and humans.

8 Limitations

One limitation of the present study is the size of the stimuli set, as a hundred problem instances, though typical in human studies, do not constitute a large-scale benchmark. Here we are limited by resources, given that human production and comprehension studies are expensive to run. However, since we have established that frontier models are generally good at our tasks, future studies may employ an LM-based augmentation to enlarge the dataset and evaluate smaller models. Furthermore, the overall scale of the human data we have collected is not small. For example, we have recruited over 700 human participants, and each comprehension problem has 40 human responses. We believe these human data would be valuable for experimental pragmatics and cognitive science.

Another limitation is our simplification of the Wavelength game. We formulate the game as single-turn comprehension and production tasks, whereas the real game features repeated interactions between multiple players. Learning about other players and adapting accordingly and quickly are implicit skills required to excel at the game, which our setup does not reflect, and such skills underlie real-world linguistic interactions. We believe richer settings where pragmatic communication, Theory-of-Mind, and personalization organically come together would be valuable for further evaluating and improving LMs as helpful, aligned conversational agents (Liu et al., 2023; Lin et al., 2024).

A limitation on the modeling side is that, although RSA is a very general computational framework, we have only explored relatively straightforward ways of combining it with LMs. For example, RSA can include considerations about costs of utterances, which might concern an utterance's length or ease of comprehension and production. Here we have made the assumption that all utterances have the same cost. More explorations of defining and utilizing utterance costs could be interesting for future LM research.

Acknowledgments

We thank Noah Goodman for general discussions about the RSA framework. We thank Guangyuan Jiang for helpful comments on the paper. We thank the three anonymous reviewers for their valuable feedback and constructive suggestions, which helped improve the quality of this work. This study was supported in part by funds from the MIT-IBM Watson AI Lab and a grant from the Simons Foundation to the Simons Center for the Social Brain at MIT. RPL acknowledges support from NSF grant BCS-2121074.

References

Kelsey Allen, Franziska Brändle, Matthew Botvinick, Judith E Fan, Samuel J Gershman, Alison Gopnik, Thomas L Griffiths, Joshua K Hartshorne, Tobias U Hauser, Mark K Ho, and 1 others. 2024. Using games to understand the mind. *Nature Human Behaviour*, pages 1–9.

Jacob Andreas and Dan Klein. 2016. Reasoning about pragmatics with neural listeners and speakers. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1173–1182, Austin, Texas. Association for Computational Linguistics.

Anthropic. 2025. Claude 3.7 sonnet and claude code.

John L Austin. 1962. *How to do things with words*. Oxford university press.

Lawrence W Barsalou. 1983. Ad hoc categories. *Memory & cognition*, 11(3):211–227.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, and 1 others. 2021. Evaluating large language models trained on code. *ArXiv preprint*, abs/2107.03374.

Gennaro Chierchia, Danny Fox, and Benjamin Spector. 2012. Scalar implicature as a grammatical phenomenon. In *Handbücher zur Sprach-und Kommunikationswissenschaft/Handbooks of Linguistics and Communication Science Semantics Volume 3*. de Gruyter.

Herbert H Clark. 1996. *Using language*. Cambridge university press.

Katherine M Collins, Ilia Sucholutsky, Umang Bhatt, Kartik Chandra, Lionel Wong, Mina Lee, Cedegao E Zhang, Tan Zhi-Xuan, Mark Ho, Vikash Mansinghka, and 1 others. 2024. Building machines that learn and think with people. *Nature human behaviour*, 8(10):1851–1863.

Max J Cresswell. 1976. The semantics of degree. In *Montague grammar*, pages 261–292. Elsevier.

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

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *ArXiv preprint*, abs/2407.21783.

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

Michael Franke. 2022. Pragmatic natural language generation with neural language models.

Daniel Fried, Nicholas Tomlin, Jennifer Hu, Roma Patel, and Aida Nematzadeh. 2023. Pragmatics in language grounding: Phenomena, tasks, and modeling approaches. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12619–12640, Singapore. Association for Computational Linguistics.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *ArXiv preprint*, abs/2312.11805.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *ArXiv preprint*, abs/2503.19786.

Noah D Goodman and Michael C Frank. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, 20(11):818–829.

Gabriel Grand, Idan Asher Blank, Francisco Pereira, and Evelina Fedorenko. 2022. Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature human behaviour*, 6(7):975–987.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Thomas L Griffiths, Nick Chater, and Joshua B Tenenbaum. 2024. *Bayesian models of cognition: Reverse engineering the mind*. MIT Press.

Irene Heim and Angelika Kratzer. 1998. *Semantics in Generative Grammar*. Blackwell, Malden, MA.

Irene Roswitha Heim. 1982. *The Semantics of Definite and Indefinite Noun Phrases*. PhD thesis, University of Massachusetts Amherst.

Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. 2016. Generating visual explanations. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 3–19. Springer.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213, Toronto, Canada. Association for Computational Linguistics.

Mingyue Jian and Siddharth Narayanaswamy. 2024. Are LLMs good pragmatic speakers? *ArXiv preprint*, abs/2411.01562.

Simeon Junker, Manar Ali, Larissa Koch, Sina Zarrieß, and Hendrik Buschmeier. 2025. Are multimodal large language models pragmatically competent listeners in simple reference resolution tasks? *ArXiv preprint*, abs/2506.11807.

Hans Kamp and Uwe Reyle. 2013. *From discourse to logic: Introduction to modeltheoretic semantics of natural language, formal logic and discourse representation theory*, volume 42. Springer Science & Business Media.

JAW Kamp. 1975. Two theories about adjectives. formal semantics of natural language, ed. by edward keenan, 123–55. In Edward L Keenan, editor, *Formal Semantics of Natural Language*, pages 123–155. Cambridge University Press.

Justine T Kao, Jean Y Wu, Leon Bergen, and Noah D Goodman. 2014. Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, 111(33):12002–12007.

Christopher Kennedy. 2007. Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and philosophy*, 30(1):1–45.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Kepa Korta and John Perry. 2024. Pragmatics. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Fall 2024 edition. Metaphysics Research Lab, Stanford University.

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. 2017. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253.

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

Stephen C Levinson. 1983. *Pragmatics*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Hengli Li, Song-Chun Zhu, and Zilong Zheng. 2023. Diplomat: A dialogue dataset for situated pragmatic reasoning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Jessy Lin, Nicholas Tomlin, Jacob Andreas, and Jason Eisner. 2024. Decision-oriented dialogue for human-ai collaboration. *Transactions of the Association for Computational Linguistics*, 12:892–911.

Benjamin Lipkin, Lionel Wong, Gabriel Grand, and Joshua B Tenenbaum. 2023. Evaluating statistical language models as pragmatic reasoners. *ArXiv preprint*, abs/2305.01020.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024a. Deepseek-v3 technical report. *ArXiv preprint*, abs/2412.19437.

Ryan Liu, Theodore R. Sumers, Ishita Dasgupta, and Thomas L. Griffiths. 2024b. How do large language models navigate conflicts between honesty and helpfulness? In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Ryan Liu, Howard Yen, Raja Marjieh, Thomas L Griffiths, and Ranjay Krishna. 2023. Improving interpersonal communication by simulating audiences with language models. *ArXiv preprint*, abs/2311.00687.

Bolei Ma, Yuting Li, Wei Zhou, Ziwei Gong, Yang Janet Liu, Katja Jasinskaja, Annemarie Friedrich, Julia Hirschberg, Frauke Kreuter, and Barbara Plank. 2025a. Pragmatics in the era of large language models: A survey on datasets, evaluation, opportunities and challenges. *ArXiv preprint*, abs/2502.12378.

Ziqiao Ma, Jing Ding, Xuejun Zhang, Dezhi Luo, Jiahe Ding, Sihan Xu, Yuchen Huang, Run Peng, and Joyce Chai. 2025b. Vision-language models are not pragmatically competent in referring expression generation. *ArXiv preprint*, abs/2504.16060.

Sonia K Murthy, Rosie Zhao, Jennifer Hu, Sham Kakade, Markus Wulfmeier, Peng Qian, and Tomer Ullman. 2025. Inside you are many wolves: Using cognitive models to interpret value trade-offs in llms. *ArXiv preprint*, abs/2506.20666.

OpenAI. 2025. Introducing GPT-4.1 in the API.

Stefan Palan and Christian Schitter. 2018. Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Isha Puri, Shivchander Sudalairaj, Guangxuan Xu, Kai Xu, and Akash Srivastava. 2025. Rollout roulette: A probabilistic inference approach to inference-time scaling of llms using particle-based monte carlo methods. *ArXiv preprint*, abs/2502.01618.

Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2023. The goldilocks of pragmatic understanding: Fine-tuning strategy matters for implicature resolution by llms. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

John R Searle. 1969. *Speech acts: An essay in the philosophy of language*, volume 626. Cambridge university press.

Dan Sperber and Deirdre Wilson. 1986. *Relevance: Communication and cognition*, volume 142. Harvard University Press Cambridge, MA.

Cesare Spinoso-Di Piano, David Eric Austin, Pablo Piantanida, and Jackie CK Cheung. 2025. $(RSA)^2$: A rhetorical-strategy-aware rational speech act framework for figurative language understanding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20898–20938, Vienna, Austria. Association for Computational Linguistics.

Settaluri Sravanthi, Meet Doshi, Pavan Tankala, Rudra Murthy, Raj Dabre, and Pushpak Bhattacharyya. 2024. PUB: A pragmatics understanding benchmark for assessing LLMs' pragmatics capabilities. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12075–12097, Bangkok, Thailand. Association for Computational Linguistics.

Robert Stalnaker. 1978. Assertion. *Syntax and Semantics (New York Academic Press)*, 9:315–332.

Michael Henry Tessler and Noah D Goodman. 2019. The language of generalization. *Psychological review*, 126(3):395.

Michael Henry Tessler and Noah D Goodman. 2022. Warm (for winter): Inferring comparison classes in communication. *Cognitive Science*, 46(3):e13095.

Michael Henry Tessler, Polina Tsvilodub, Jesse Snedeker, and Roger P Levy. 2020. Informational goals, sentence structure, and comparison class inference. In *Proceedings of the Annual Conference of the Cognitive Science Society*.

Polina Tsvilodub, Kanishk Gandhi, Haoran Zhao, Jan-Philipp Fränken, Michael Franke, and Noah D Goodman. 2025. Non-literal understanding of number words by language models. *ArXiv preprint*, abs/2502.06204.

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

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. 2024. Self-preference bias in llm-as-a-judge. *ArXiv preprint*, abs/2410.21819.

Ludwig Wittgenstein. 1953. *Philosophical Investigations*. Wiley-Blackwell, New York, NY, USA.

Lionel Wong, Gabriel Grand, Alexander K Lew, Noah D Goodman, Vikash K Mansinghka, Jacob Andreas, and Joshua B Tenenbaum. 2023. From word models to world models: Translating from natural language to the probabilistic language of thought. *ArXiv preprint*, abs/2306.12672.

Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. 2023. A brief overview of chatgpt: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, 10(5):1122–1136.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, and 1 others. 2025. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2):121101.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *ArXiv preprint*, abs/2505.09388.

Lance Ying, Katherine M Collins, Lionel Wong, Ilia Sucholutsky, Ryan Liu, Adrian Weller, Tianmin Shu, Thomas L Griffiths, and Joshua B Tenenbaum. 2025. On benchmarking human-like intelligence in machines. *ArXiv preprint*, abs/2502.20502.

Erica J Yoon, Michael Henry Tessler, Noah D Goodman, and Michael C Frank. 2020. Polite speech emerges from competing social goals. *Open Mind*, 4:71–87.

Haoran Zhao and Robert D Hawkins. 2025. Comparing human and llm politeness strategies in free production. *ArXiv preprint*, abs/2506.09391.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Zilong Zheng, Shuwen Qiu, Lifeng Fan, Yixin Zhu, and Song-Chun Zhu. 2021. GRICE: A grammar-based dataset for recovering implicature and conversational rEasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2074–2085, Online. Association for Computational Linguistics.

A Human Experiment Details

Here we include more details of human data collection.

Procedure. As described in Section 3, the set of 100 problems constitutes the stimuli for our production data collection. Each participant generates a clue for a given problem, and they do so for 10 problems. To ensure data quality we ask the participant to think for at least 20 seconds before generating the clue. We collect 5 clues for each problem and 500 clues in total. Participants' median completion time for the generation task is approximately 9 minutes. To potentially increase clue diversity and quality, the author team also composes one clue for each problem. We combine these clues with participant-composed clues for the next stage—filtering out the best clue for each problem, for which we collect 15 human judgments (guessing the target value given the corresponding clue using a slider). We keep the best clues based on mean human absolute differences from the target value, finalizing a set of 100 problems for the official comprehension task.

In the comprehension task, similarly to ensure data quality we ask the participant to think for at least 10 seconds before guessing the target value using a slider. Each participant makes guesses for approximately 20 problems, which has a median completion time of 8 minutes. On each comprehension problem, we collect 25 judgments, and plus the previous 15 judgments we have total 40 human judgments. This relatively large number of judgments allows us to have a more representative distribution of human uncertain judgments.

Instructions and interface. The main interface for the human experiments is shown in Fig. 7. The verbal instructions we provide to human participants are highly similar to the listener and speaker prompts we provide to LMs (shown below in Section F).

Cost. Human participants on average are paid at a rate of at least \$15 per hour. The production (clue generation) human experiment costs ∼\$120. The comprehension human experiment (including filtering for good clues) costs ∼\$1000. The human evaluation of model generated clues costs ∼\$150. So the overall cost is ∼\$1270.

B Language Model Details

All open-weights models except Deekseek-V3 are run on local H100 and A100 clusters, and all corresponding experiments can be finished in 12 hours with 4 H100s. Deepseek-V3 is run on the

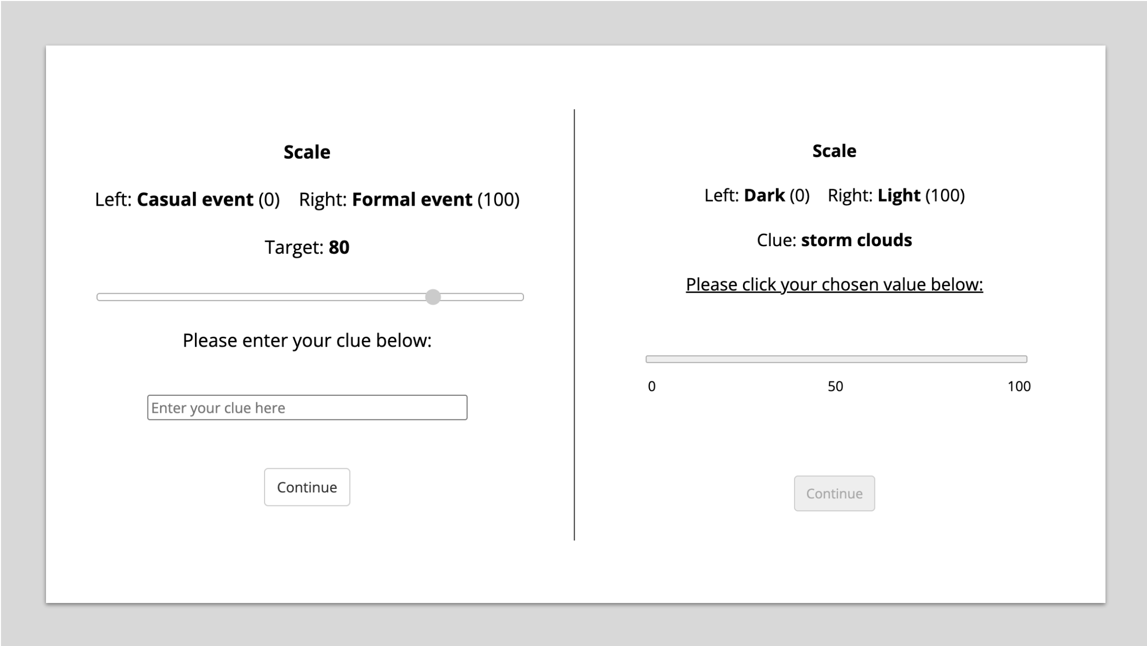

Figure 7: Experimental interface for the human studies. One the left and right show the production (clue generation) and comprehension (target guessing) interfaces, respectively.

Together AI inference platform via APIs. Close models Gemini 2.0 Flash (`001`), Claude 3.7 Sonnet (`20250219`), and GPT-4.1 (`2025-04-14`) are called using their respective APIs. All API calls combined cost under $500.

C Alternative RSA Listener Model

We explore an alternative RSA model for the LM comprehension task. Specifically, we define the pragmatic listener L_1 as

$$P_{L_1}(s \mid u) = \frac{P_{L_0}(s \mid u) \cdot P_{S_0}(u \mid s)}{\sum_{s' \in \mathcal{S}} P_{L_0}(s' \mid u) \cdot P_{S_0}(u \mid s')}.$$

This model is similar to our pragmatic speaker S_1 in Section 5, except that we marginalize over all states $\mathcal{S}$ instead of all utterances $\mathcal{U}$. We use the length-normalized probability of an directly prompted LM as an unnormalized estimate of $P_{S_0}(u \mid s)$. Since this requires having access to the model's next-token probability, we only evaluate this using open-weights models.

We show the results in Fig. 8 and Fig. 9. We find that this alternative performs similarly to the RSA model described in Section 4, and incorporating RSA does not necessarily improve task performance or human-likeness on the comprehension task.

D Additional LM Results

D.1 Listener Distribution Entropy

We measure the entropy of the listener distribution and show results in Fig. 10. We find that the LM distributions generally have lower entropy than that of human distributions, indicating that human distributions reflect more uncertainty. We observe that RSA almost always increases entropy over the base prompting method.

D.2 Comprehension Task Performance Breakdown

While our target values uniformly range from 0 to 100, some might be easier to guess. In Fig. 11, we show the average absolute difference broken down by target value. We find that the smaller LM (Qwen3 8B) has larger variance across target values, whereas humans and the larger LM (Qwen3 32B) are more consistent.

D.3 Choices of Alternatives in the Comprehension Task

We do not observe improvements when incorporating the RSA model for the comprehension task. Here, we investigate two possible hypotheses for why it does not provide further improvement: (1) the LM listener distribution is concentrated and spiky, which limits the benefits of RSA, or (2) the LM does not generate good alternatives as a

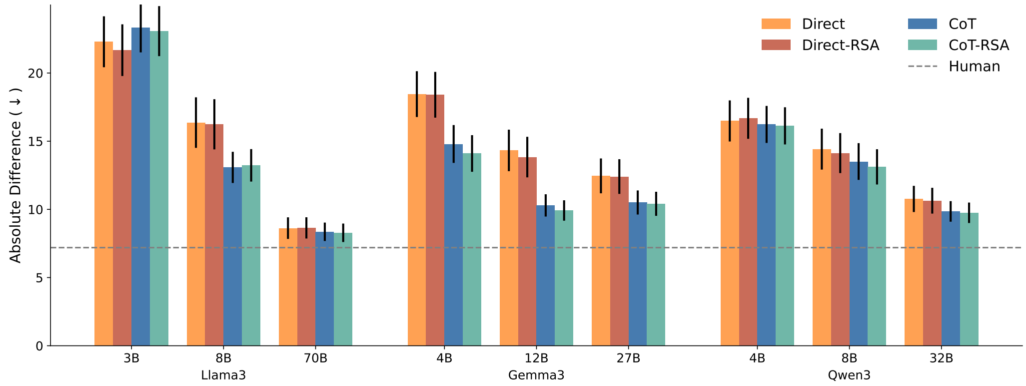

Figure 8: Absolute difference between the model's prediction and the target value. We use the alternative RSA model for Direct-RSA and CoT-RSA. Error bars show standard error over each problem. The dashed line indicates human performance.

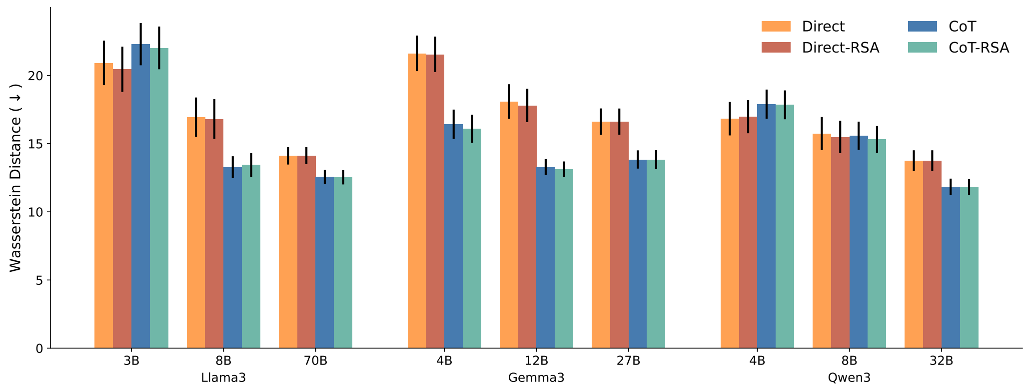

Figure 9: The Wasserstein distance between the model distribution and human distribution. We use the alternative RSA model for Direct-RSA and CoT-RSA. Error bars show standard error over each problem.

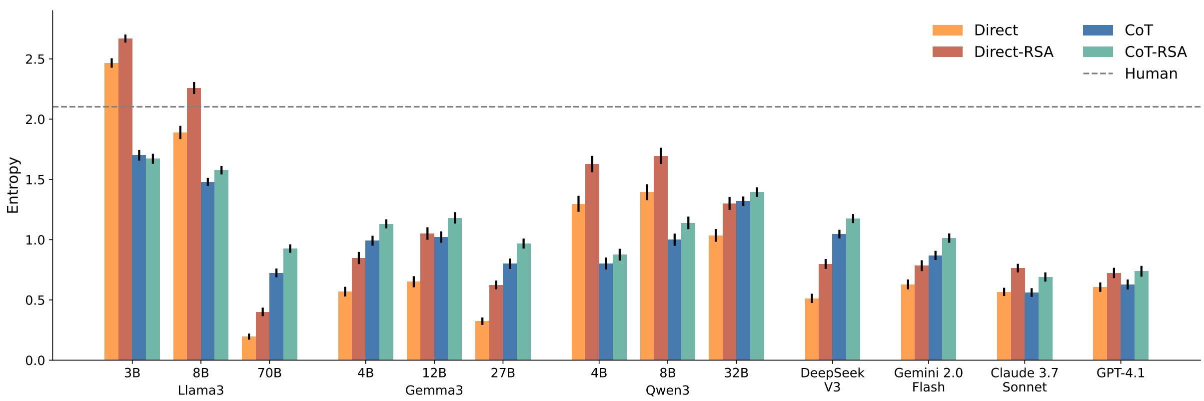

Figure 10: Entropy of the LM listener distribution. Error bars show standard error over each problem. The dashed line indicates human performance.

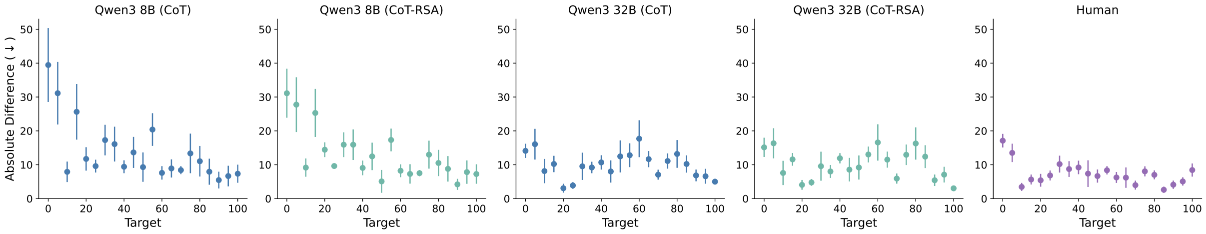

Figure 11: The average absolute difference between the model or human prediction and the target, broken down by target value. Humans perform worse at the extreme scale (0), but are generally consistent across different target values. The performance of LMs, however, shows larger variance across target values, especially for small model (Qwen3 8B) which performs significantly poorly near 0. Error bars show standard error over each target value.

			Absolute Difference (↓)		Wasserstein Distance (↓)	
Model	Listener	Speaker	Direct-RSA	CoT-RSA	Direct-RSA	CoT-RSA
Llama3	3B	3B	22.49	21.63	20.71	21.13
	3B	70B	22.05	21.79	20.39	20.69
	8B	8B	15.49	12.89	15.44	12.30
	8B	70B	16.16	12.96	16.01	12.65
Gemma3	4B	4B	18.25	14.48	20.90	16.25
	4B	27B	17.84	14.49	20.35	16.10
	12B	12B	13.77	10.35	17.13	12.88
	12B	27B	13.46	10.45	16.49	12.89
	27B	4B	12.14	10.06	16.09	13.40
	27B	12B	12.02	10.82	15.98	13.71
	27B	27B	12.07	10.51	15.87	13.11
Qwen3	4B	4B	17.50	16.70	17.42	18.22
	4B	32B	16.97	15.76	16.91	17.54
	8B	8B	13.37	12.65	14.36	14.22
	8B	32B	13.82	13.19	14.50	15.02
	32B	4B	11.30	9.93	13.53	11.79
	32B	8B	10.82	9.95	13.24	11.94
	32B	32B	11.00	10.05	13.25	12.02

Table 4: Absolute difference between the model's prediction and the target value, and the Wasserstein distance between the model distribution and the human distribution. We compare the model's performance under two conditions: using alternatives generated by the model itself versus those generated by a different LM.

simulated speaker.

To investigate which factor contributes more to the performance, we provide each listener LM with alternatives generated from a different speaker LM. Since the stronger LMs tend to generate better alternatives (as evidenced by their better performance on the generation task in Fig. 6), if the weaker model benefits from better alternatives, this would suggest its performance is limited by its ability to generate good alternatives. Similarly, if a stronger model's performance degrades when using alternatives from a weaker LM, it would confirm that the ability to propose good alternatives is a key factor. In contrast, if there is no significant performance difference when an LM uses its own alternatives versus those from another model, this would suggest that the LM's listener distribution may be the main issue.

We show results in Table 4. For most models, we do not observe a significant difference between the original results and the ones using alternatives from another model, which suggests that for the comprehension task, the performance is less sensitive to the choices of alternatives. Therefore, we hypothesize that the LM's listener distribution may be a more important limiting factor.

D.4 Number of Alternatives in the Production Task

In the LM production task setting, we observe significant improvement using RSA on top of both direct and CoT prompting. One important component of RSA is reasoning about how a listener will interpret the meaning of alternative utterances. To investigate the impacts of the number of utterances, we evaluate the speaker models but use different

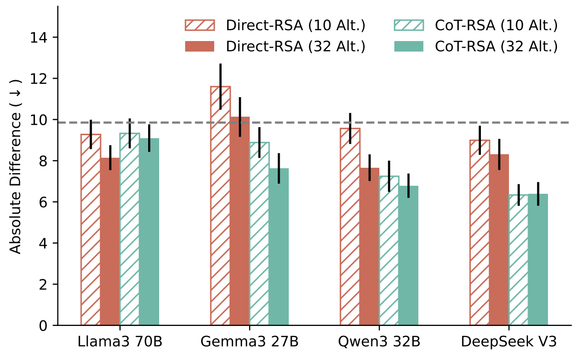

Figure 12: Ablations of the number of alternatives for the language production task. We compare the RSA-based LM speakers using 10 or 32 alternatives. The results show the absolute difference between the listener model's (Qwen3 32B) prediction and the target value.

numbers of utterances. As shown in Fig. 12, across the best-performing models from different families, increasing the number of alternatives consistently improves performance.

D.5 Full Results

We show the correlations between model predictions and human judgments in Fig. 13–Fig. 16.

E Qualitative Example Outputs

The following is an example output of Qwen3 32B listener using CoT prompting.

Qwen3 32B listener output using CoT

```
To determine where "soup" falls on the scale between
Snack (0) and Meal (100), we can consider whether soup
is generally associated more with a light snack or a
full meal.

- **Snack (0):** Snacks are typically small, quick, and
meant to satisfy hunger briefly. Examples include
chips, fruits, or a granola bar.
- **Meal (100):** Meals are more substantial, often
part of a regular eating pattern (like breakfast,
lunch, or dinner), and may include multiple components.

- **Soup as a clue:** Soup can be both a snack and part
of a meal, depending on the context. For example, a
small can of soup eaten on the go might be a snack, but
a bowl of soup served with a salad and bread could be
part of a full lunch. Soup is often served as an
appetizer or a main dish in a meal.

Given this flexibility, but leaning toward being a
component of a meal rather than a simple snack, a
reasonable estimate for "soup" on the Snack-Meal scale
would fall around the **middle to slightly higher**
part of the spectrum.

<answer>60</answer>
```

The following is an example output of Qwen3 32B speaker using CoT prompting.

Qwen3 32B speaker output using CoT

```
The target value is 70 on a scale from Colorless (0) to
Colorful (100), which means the clue should represent
something that is **fairly colorful**, but not
**extremely** vibrant or multicolored. It should be on
the **brighter end** of the spectrum, but still allow
room for something more intensely colorful to exist. A
strong clue would evoke something that has **distinct
color**, but may not have **a wide variety of colors**.

Good clues might include:
- **Sunset** - vibrant but not multicolored in the
extreme
- **Tropical fish** - colorful but still short of the
most dazzling hues
- **Rainbow flags** - colorful but not the most vivid
version of color (e.g., lasers, stained glass)
- **Flowers** - can be colorful, but the term is
general enough to allow for variation
- **Marigold** - a specific, strong color but still
within the colorful range

Clues that would be **too colorless** might include:
- Gray (closer to 0)
- Concrete (very near to 0)

Clues that would be **too colorful** might include:
- Neon (closer to 100)
- Stained glass (also closer to 100)

Therefore, a well-chosen clue for 70 might be:

<answer>sunrise</answer>
```

F Prompt Details

The following are our prompts for eliciting listener judgments and speaker productions from LMs.

Listener prompt

```
# Context

You will be playing a language game. In each round, you
will be presented a scale from 0 (left) to 100 (right).
Each scale represents a spectrum between two opposite
or contrasting concepts. One concept will be on the
left side of the scale (towards 0), and the other
concept will be on the right side (towards 100).

For each scale, you will see a clue-a single word or a
short phrase-that someone else has provided. Your task
is to guess a value between 0 and 100 (in increments of
5) that you believe best represents where that clue
falls on the scale between the two concepts.
Essentially, you are trying to interpret where on the
spectrum the clue giver intended their clue to
represent. A good guess should be close to the target
value that the clue giver was trying to communicate.
(The clue giver chose the clue with respect to the
value provided to them.)

Let's look at an example:

Imagine the scale is from Slow (0) to Fast (100).

If the clue is 'race car', you would likely guess a
high value, perhaps around 85, since a race car
represents something that is very fast, but not the
absolute fastest thing possible. If the clue had been
'turtle' instead, you might guess a value closer to 10,
or if it had been 'bicycle', you might guess a value
closer to 50 on this particular scale.

Here's another example:

Scale: Summer (0) to Winter (100).
Clue: ice cream.

This clue is more strongly associated with 'Summer'
than with 'Winter'. A reasonable guess might be around
30. While people certainly eat ice cream year-round, it
is especially popular during hot summer days. If the
clue had been 'wearing a shirt' you might guess a value
closer to 50, or if it had been 'snow' you might guess
a value closer to 90 or 100 in this context.

Here're a few more examples:

Scale: Feels bad (0) to Feels good (100).
Clue: relaxing.
<answer>70</answer>

Scale: Hard to spell (0) to Easy to spell (100).
Clue: Daenerys Targaryen.
<answer>20</answer>

Scale: Red (0) to Yellow (100).
Clue: tangerine.
<answer>50</answer>

Provide your best estimate carefully (in increments of
5, so the possible values are 0, 5, 10, 15, 20, 25, 30,
35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95,
100). The goal is to get as close as possible to the
target value that the clue giver intended to
communicate. If you find a clue confusing, meaningless,
or hard to understand, please still make an educated
guess. Some clues may not be very good, but please
always make an honest attempt.

Format your response as:

<answer>your guess</answer>

# Problem

Scale: {left_word} (0) to {right_word} (100).
Clue: {clue}.
```

Speaker prompt

```
# Context

You will be playing a language game. In each round, you
will be presented a scale from 0 (left) to 100 (right).
Each scale represents a spectrum between two opposite
or contrasting concepts. One concept will be on the
left side of the scale (towards 0), and the other
concept will be on the right side (towards 100).

For each scale, you will see a target value, indicated
by a number between 0 and 100. Your task is to think of
a clue (often times a single word, or a very short
phrase if necessary) that you believe best represents
that specific target value's position on the scale,
considering the two concepts. Essentially, you are
trying to communicate where the target value lies on
the spectrum using just your clue. A good clue should
allow another person to approximately guess what the
target value is.

Let's look at an example:

Imagine the scale is from Slow (0) to Fast (100).

If the target value is 85, you need a clue that
represents something that is very fast, but perhaps not
the absolute fastest. Good potential clues might
include: 'professional sprinters', 'leopard', or
'Ferarri'. A clue like 'turtle' (which might be closer
to 0) or 'bicycle' (which might be closer to 50) would
be less appropriate for the target of 85 on this
particular scale.

Here's another example:

Scale: Summer (0) to Winter (100).

Target value: 30. This target is closer to 'Summer'
than to 'Winter'. Good clues might be: 'swimming', 'ice
cream', or 'sunscreen'. Clues like 'wearing a shirt'
(arguably closer to 50) or 'snow' (arguably closer to
100) would be less fitting for a target of 30 in this
context.

Here're a few more examples:

Scale: Feels bad (0) to Feels good (100).
Target value: 70.
<answer>relaxing</answer>

Scale: Hard to spell (0) to Easy to spell (100).
Target value: 20.
<answer>Daenerys Targaryen</answer>

Scale: Red (0) to Yellow (100).
Target value: 50.
<answer>tangerine</answer>

You are encouraged to be concise when you come up with
the clue. Try using a single word or a short phrase (a
few words). You are not allowed to use more than 5
words for a clue. Your clue should convey a single
thought. Do not combine multiple ideas. Do not use
words that already appear in the two concepts or any
synonyms (e.g., 'Piece' is not allowed for the
'Peaceful - Warlike' scale.) Do not use modifiers
(words like 'but', 'very', 'almost', and 'slightly').
Do not use specific numbers or numeric values
(including time, percentage, etc.). Do not use
modifiers (words like 'but', 'very', 'almost', and
'slightly'). Do not use specific numbers or numeric
values (including time, percentage, etc.).

Provide the best clue you can think of. The best clues
maximize the chance that an average person can
approximately guess the target value given the clue.

Format your response as:

<answer>your clue</answer>

# Problem

Scale: {left_prompt} (0) to {right_prompt} (100).
Target value: {target_value}.
```

G Concept Pairs

Here is a table presenting all the concept pairs used as stimuli in our study.

Index	Left Concept	Right Concept
1	Bad	Good
2	Hot	Cold
3	Colorless	Colorful
4	Low calorie	High calorie
5	Inessential	Essential
6	Cheap	Expensive
7	Rare	Common
8	Difficult to use	Easy to use
9	Worst day of the year	Best day of the year
10	Bad habit	Good habit
11	Dark	Light
12	Hard to remember	Easy to remember
13	Unhealthy	Healthy
14	Normal pet	Exotic pet
15	Happens slowly	Happens suddenly
16	Mental activity	Physical activity
17	Need	Want
18	Dry food	Wet food
19	Optional	Mandatory
20	Hard to pronounce	Easy to pronounce
21	Low quality	High quality
22	Plain	Fancy
23	Quiet place	Loud place
24	Dangerous	Safe
25	Useless major	Useful major
26	Bad for you	Good for you
27	Waste of time	Good use of time
28	Nobody does it	Everybody does it
29	Snack	Meal
30	Soft	Hard
31	Square	Round
32	Temporary	Permanent
33	Sport	Game
34	Messy food	Clean food
35	Vice	Virtue
36	Unpopular activity	Popular activity
37	Boring	Exciting
38	Easy to do	Hard to do
39	Nature	Nurture
40	Limited	Infinite
41	Casual event	Formal event
42	Small talk	Heavy topic
43	Short	Long
44	Talent	Skill
45	Unnatural	Natural
46	Funny topic	Serious topic
47	Not enough	Too much
48	Art	Commerce
49	Deep thought	Shallow thought
50	Blue	Green

Table 5: All fifty pairs of concepts used as the experimental stimuli.

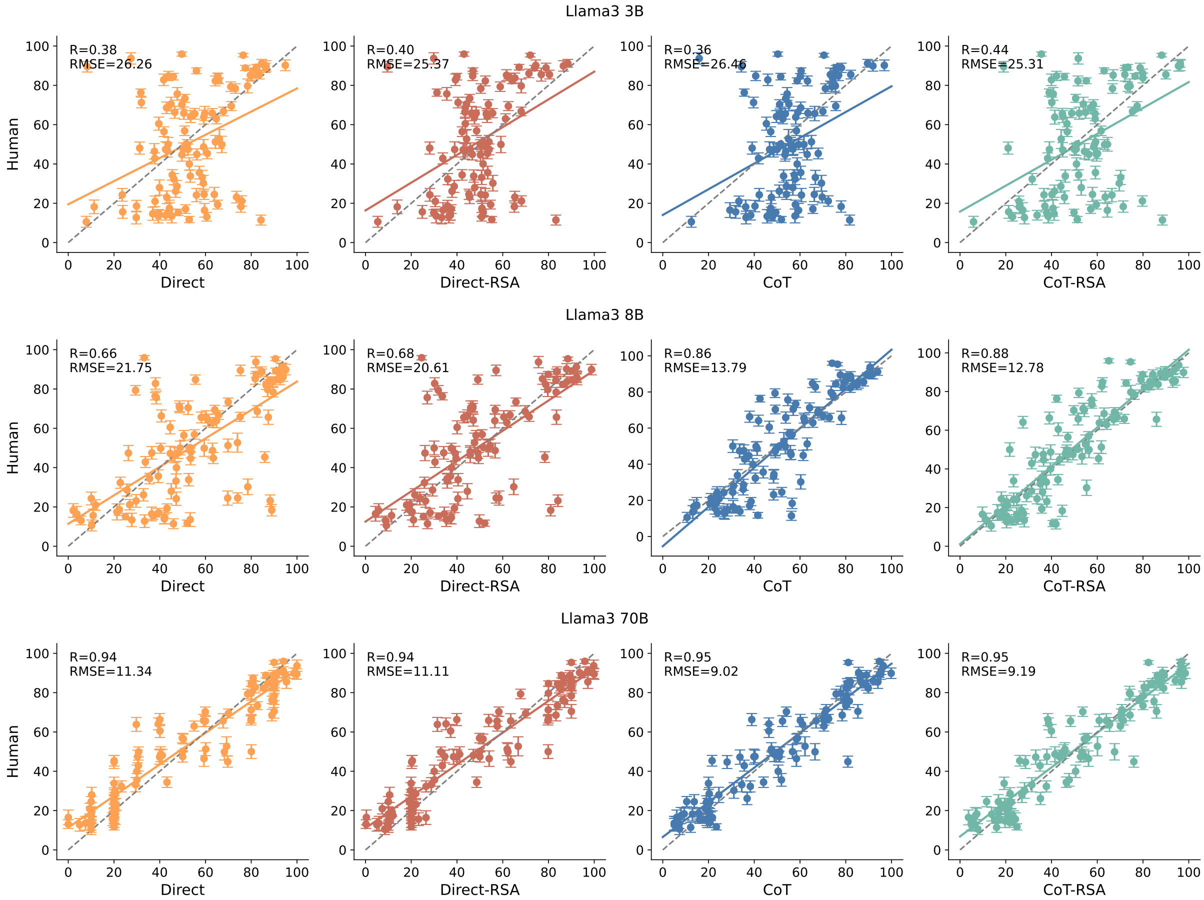

Figure 13: Correlations between model predictions and human judgments using Llama3 models. We show Pearson correlations and root mean square standard error (RMSE). Error bars show standard error over 40 human participants.

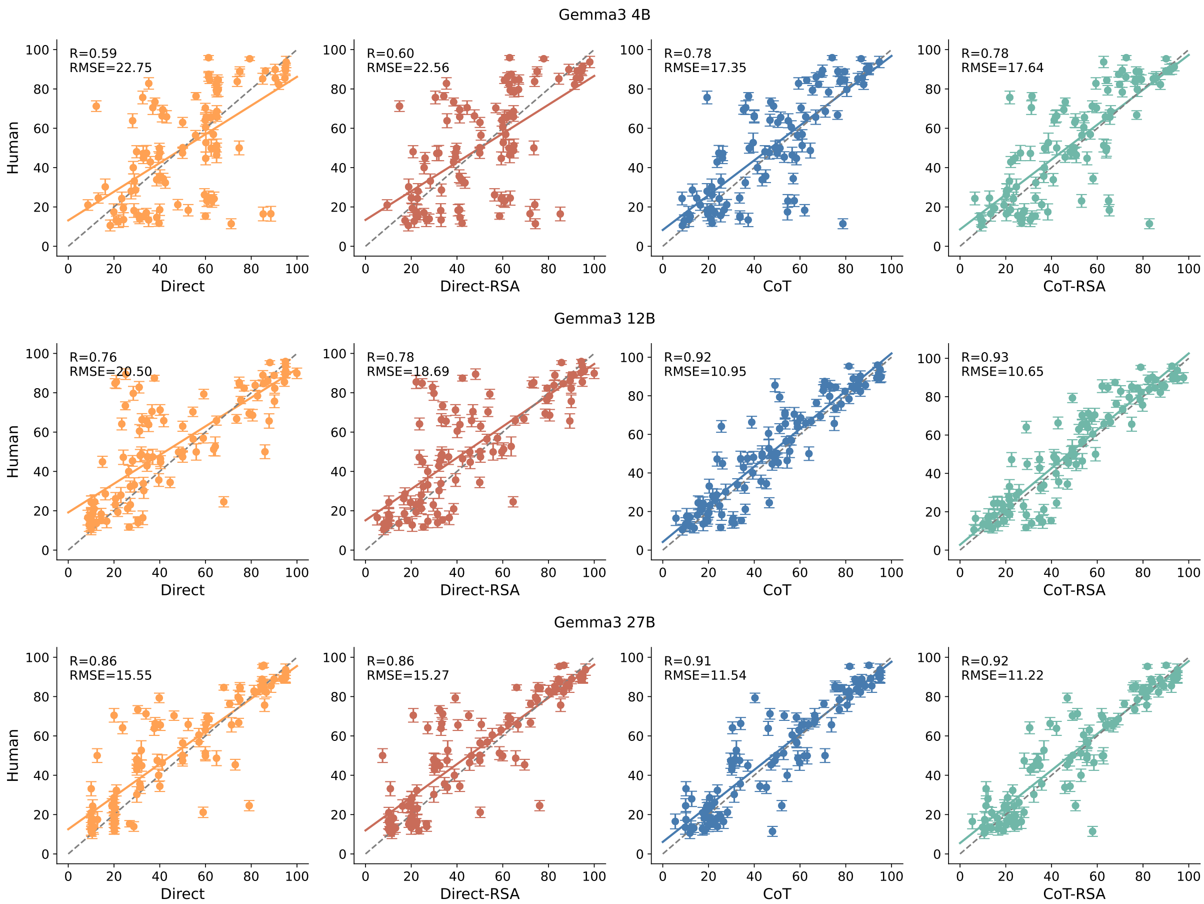

Figure 14: Correlations between model predictions and human judgments using Gemma3 models. We show Pearson correlations and root mean square standard error (RMSE). Error bars show standard error over 40 human participants.

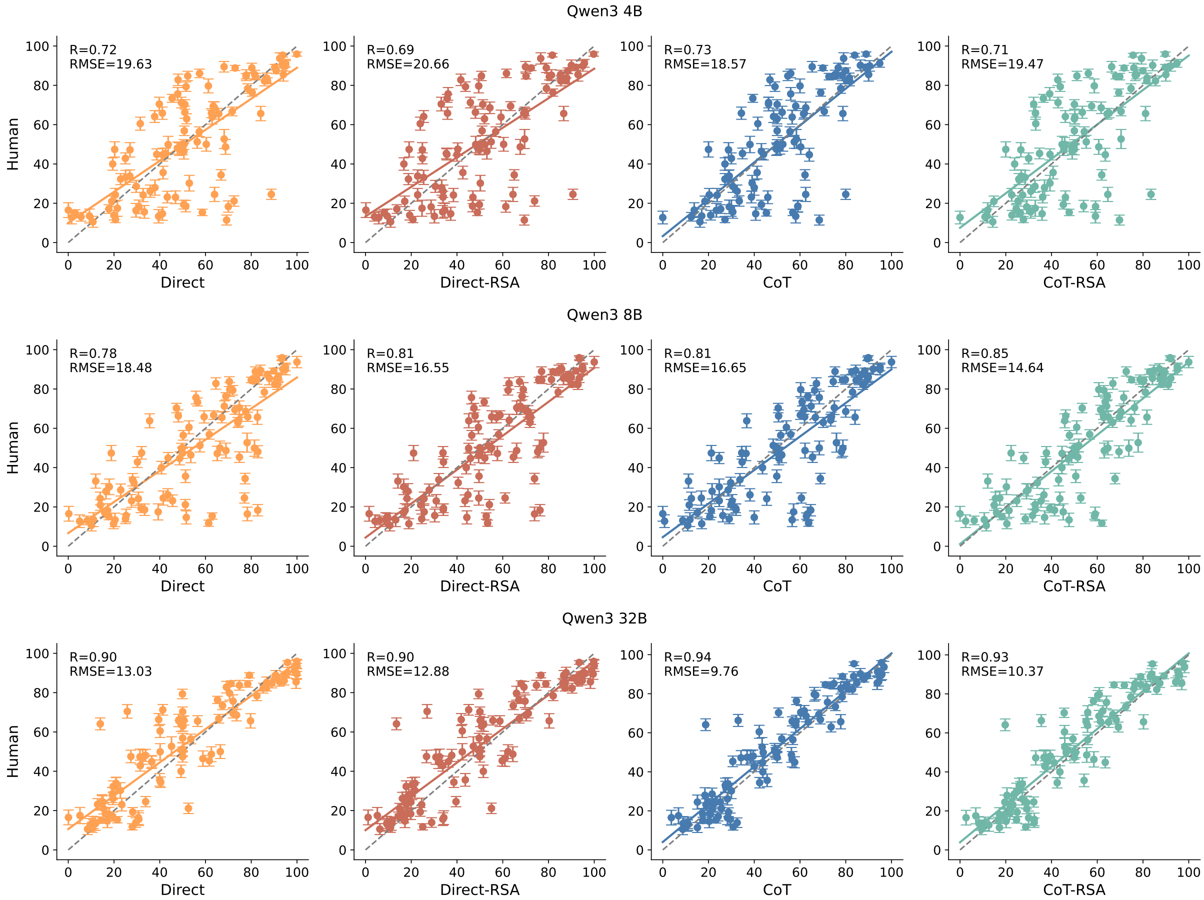

Figure 15: Correlations between model predictions and human judgments using Qwen3 models. We show Pearson correlations and root mean square standard error (RMSE). Error bars show standard error over 40 human participants.

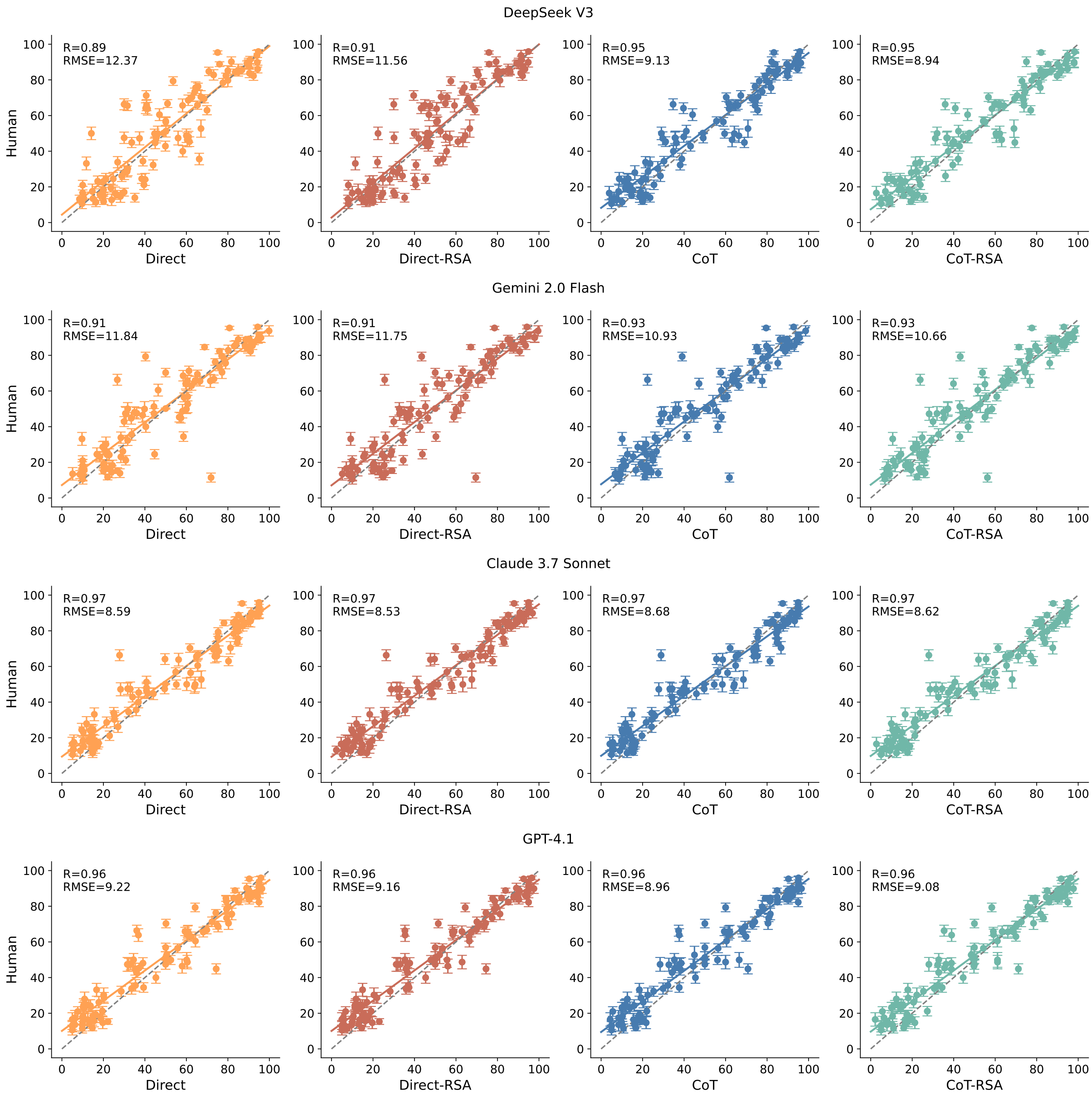

Figure 16: Correlations between model predictions and human judgments using DeepSeek-V3 and closed-source models. We show Pearson correlations and root mean square standard error (RMSE). Error bars show standard error over 40 human participants.