Humans and LLMs Diverge on Probabilistic Inferences

Gaurav Kamath[α,β] **Sreenath Madathil**[α,β] **Sebastian Schuster**[γ]
Marie-Catherine de Marneffe[τ] **Siva Reddy**[α,β,δ]
[α] McGill University [β]Mila – Quebec AI Institute
[γ]University of Vienna [τ]FNRS – UCLouvain [δ]Canada CIFAR AI Chair

Abstract

Human reasoning often involves working over limited information to arrive at probabilistic conclusions. In its simplest form, this involves making an inference that is not strictly entailed by a premise, but rather only likely given the premise. While reasoning LLMs have demonstrated strong performance on logical and mathematical tasks, their behavior on such open-ended, non-deterministic inferences remains largely unexplored. We introduce PROBCOPA, a dataset of 210 handcrafted probabilistic inferences in English, each annotated for inference likelihood by 25–30 human participants. We find that human responses are graded and varied, revealing probabilistic judgments of the inferences in our dataset. Comparing these judgments with responses from eight state-of-the-art reasoning LLMs, we show that models consistently fail to produce human-like distributions. Finally, analyzing LLM reasoning chains, we find evidence of a common reasoning pattern used to evaluate such inferences. Our findings reveal persistent differences between humans and LLMs, and underscore the need to evaluate reasoning beyond deterministic settings.[1]

1 Introduction

Much of the day-to-day reasoning that humans do involves working over partial information to arrive at probabilistic conclusions (Oaksford and Chater, 2007). Consider:

(1) *There was an accident on the highway.* → *Traffic was worse than usual.*

(2) *There was an accident on the highway.* → *Traffic was largely unaffected.*

In the absence of any further context, (1) and (2) involve conclusions that may or may not be true given the information presented. Instead, the two conclusions are only likely or unlikely to varying degrees given the context. Although (1) is likely, it is not guaranteed—perhaps everyone avoided the highway after hearing the news, leading to less traffic. Conversely, although (2) is unlikely, it cannot be ruled out completely—maybe the vehicles involved were swiftly moved out of the way, leading to minimal impact on traffic. We refer to such reasoning as *probabilistic reasoning*, and individual inferences of the kind in (1) and (2) as *probabilistic inferences*.

What does such reasoning look like in humans and large language models (LLMs)? In this paper, we provide initial insights into this question. We compare probabilistic reasoning in humans and LLMs, in terms of their respective judgments towards a range of commonsense probabilistic inferences (see Figure 1). Overall, we find that while models generally align with human judgments for probabilistic inferences deemed highly likely or highly unlikely, they consistently struggle to align with human judgments towards probabilistic inferences where annotators show more uncertainty (i.e., inferences that were deemed neither highly unlikely nor highly likely) and almost never show human-level judgment variation across sampled responses. We make the following contributions:

- We introduce PROBCOPA, a novel dataset of 210 handcrafted probabilistic inferences in English, with at least 25 human annotations per item.
- We highlight persistent differences between humans and LLMs in their judgments towards such probabilistic inferences.
- We identify patterns in LLM reasoning chains that shed light on how they arrive at their final responses in these contexts.

[1]All data and code can be found at github.com/McGill-NLP/probabilistic-reasoning

Transactions of the Association for Computational Linguistics, vol. 14, pp. 2057–2080, 2026. https://doi.org/10.1162/TACL.a.791
Action Editor: Afra Alishahi. Submission batch: 1/2026; Published 8/2026.
©2026 Association for Computational Linguistics. Distributed under a CC-BY 4.0 license.

Downloaded from direct.mit.edu/tacl/article-pdf/doi/10.1162/TACL.a.791/2624521/tacl.a.791.pdf by guest on 20 September 2026

2 The PROBCOPA Dataset

2.1 Data Construction

We aim to study inferences that are not strictly logically entailed, but rather those that lie on a range of likelihood given a premise. Due to the limitations of existing NLI datasets in this regard (see Section 7), we construct our own corpus of probabilistic inferences.

We begin with the Choice of Plausible Alternatives (COPA) dataset (Roemmele et al., 2011), which consists of 1,000 manually handcrafted items that probe commonsense reasoning.

(3) Premise: A drought occurred in the region.
What happened as a result?
   a. Alternative 1: The crops perished.
   b. Alternative 2: The water became contaminated.

As the example in (3) illustrates, each COPA item consists of a single premise together with two possible effects or causes. For each pair of alternatives, one alternative is more plausible than the other; accordingly, the original task formulation involves choosing the *more likely* alternative among the two ((3a) in this example). Importantly, however, both alternatives are designed to be at least slightly plausible given the premise.

To construct our dataset, we therefore split each COPA item into two NLI-style items, such as (4) and (5):

(4) P: A drought occurred in the region.
H: The crops perished.

(5) P: A drought occurred in the region.
H: The water became contaminated.

Crucially, although the alternatives are designed to yield clear preferences when evaluated against each other, when each is evaluated in isolation, they constitute probabilistic inferences with varying degrees of plausibility or likelihood.

Due to frequently-attested complexities in people's estimation of causal likelihood (Eddy, 1982; Villejoubert and Mandel, 2002; Krynski and Tenenbaum, 2007; Stilgenbauer et al., 2017), we exclude COPA items that ask participants to reason over possible causes, and only include those that elicit judgments on effect likelihood given a premise. We take a random sample of 105 such items from the COPA test set and split each of these as described above, resulting in 210 probabilistic inferences framed as NLI-style items.

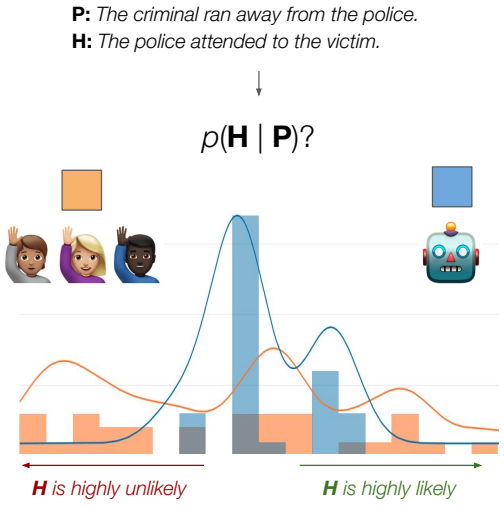


Figure 1: High-level overview of our paper. We use PROBCOPA, a novel dataset of probabilistic inferences, to collect judgments of inference likelihood from humans and models, and study how well these align with one another.

2.2 Human Annotation Procedure

We conducted online crowdsourced experiments via Prolific[2] to obtain human annotations for our dataset. We recruited 328 native English speakers based in the U.K., U.S. or Canada; these participants each annotated up to 30 PROBCOPA items under the procedure described below. All experimental protocols with humans were approved by our institution's Research Ethics Board, and participants were paid an average of US$15.00/hr.

The annotation procedure involved crowdworkers being presented with one premise-hypothesis pair at a time, and rating the likelihood of the hypothesis as a result of the premise (using a sliding scale to return a numerical rating between 0 and 100). Given attested variation in how humans express likelihood and uncertainty (Change et al., 2007; Wintle et al., 2019; Ulmer et al., 2025), the sliding scale was shown to participants along with an aid suggesting how to partition values along it.

Participants began with five instructional examples for which they received feedback—this was meant to both explain the task format to them, as well as calibrate their responses within broad ranges of the numerical scale. Following these examples, participants were presented with a sample of up to 30 test stimuli, with five attention checks interspersed in between. The order of test stimuli

[2] https://www.prolific.com

Downloaded from direct.mit.edu/tacl/article-pdf/doi/10.1162/TACL.a.791/2624521/tacl.a.791.pdf by guest on 20 September 2026

was randomly shuffled for each participant, and all responses from participants who failed more than one attention check were discarded.

After discarding data from participants who failed the attention checks, we were left with between 25-30 likelihood score annotations (each from a unique participant) for each of our 210 items, with a median of 28 annotations per item.

Appendix A further details the human annotation set-up, including screenshots of its user interface.

2.3 Reproducibility of Human Responses

We run two rounds of validation to ensure that our human responses are reproducible. In the first, reported further in Section 4.1, we have 30 items re-annotated by 30 new participants, and use this as a baseline for human-to-human response comparisons. In the second round, we have the same 30 items re-annotated by another 30 new participants, but this time with a slightly different prompt wording. We calculate the Spearman correlation between mean item ratings from our original annotations and each of these validation rounds; Spearman's $\rho = 0.98$ ($p = 4.52e-20$) and 0.97 ($p = 1.22e-19$) for the first and second validation rounds respectively. Similarly, using two-sample Kolmogorov-Smirnov tests, we find no statistically significant differences in human response distributions under either of these conditions ($\alpha = 0.05$). Together, these validation results suggest our human annotations are strongly reproducible and trustworthy.

3 Analysis of Human Responses

3.1 Methodology

On Normalizing Human Responses Studies that analyze human responses on a numerical scale often normalize human ratings (typically via by-participant z-scoring) to allow for comparisons across participants who may use the scale differently (e.g., Sprouse et al., 2013; Mahowald et al., 2016; Pavlick and Kwiatkowski, 2019). In this study, however, we deliberately instead work with the raw likelihood scores from participants, and analyze their distribution in relation to factors like human response time or their entropy.

We do so primarily for the following reason: since our scale explicitly corresponds to event likelihood, we often actually *want* to preserve inter-annotator differences in scale use. For example, if an annotator only responds with values between 0 and 95, we argue that this means the annotator specifically chooses to never assign full likelihood (100) to an event, and that this behavioral pattern should not be normalized away. Moreover, given that participants received guidance and feedback on scale use prior to and during annotation, and were subject to attention checks (see Section 2.2, Appendix A), we trust that their responses reflect their own likelihood judgments, rather than merely being artifacts of their scale use.

Metric for Response Spread To quantify how spread out human responses are for each item, we use *differential entropy*, an extension of Shannon entropy to continuous variables. For a continuous random variable X with probability density function $f(x)$, the differential entropy is defined as:

$$h(X) = -\int f(x) \log f(x) dx \quad (6)$$

Higher differential entropy values indicate greater dispersion in responses, while lower values indicate more concentration. We employ differential entropy over other metrics of spread (such as variance) due to its mathematical properties—intuitively, it captures the spread of information, rather than simply distance from the mean. For instance, while a bimodal distribution with responses concentrated at either extreme of our scale would yield extremely high variance, its differential entropy would not be as high, as the information remains relatively tightly clustered (even if this is into two groups). Crucially, however, unlike Shannon entropy, differential entropy can take negative values when a distribution is extremely concentrated, as we see for a handful of items in Figure 4.

3.2 Results

Likelihood scores from humans reveal graded, probabilistic judgments. Figure 2 (*top-left*) shows the overall distribution of likelihood scores from human annotators, across the whole PROBCOPA dataset. As it indicates, while the distribution of likelihood scores across the entire dataset shows three clear modes—corresponding to very low, very high, and balanced inference likelihood—a significant proportion of likelihood scores provided by annotators lies in between these modes, corresponding to more graded likelihood judgments.

In Appendix E, we show how this compares to similar human response data collected by Pavlick and Kwiatkowski (2019) on several major NLI

Downloaded from direct.mit.edu/tacl/article-pdf/doi/10.1162/TACL.a.791/2624521/tacl.a.791.pdf by guest on 20 September 2026

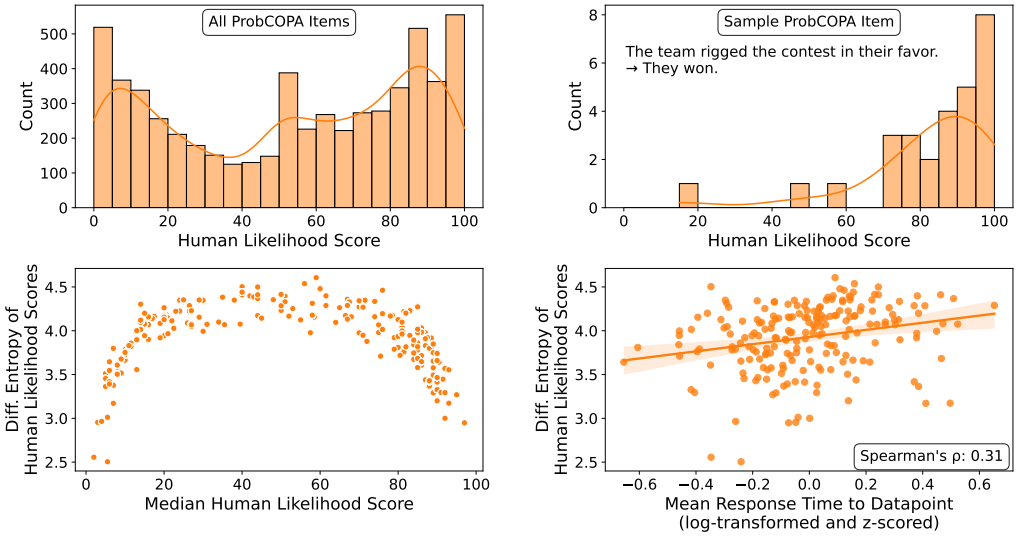


Figure 2: Distribution of human responses to PROBCOPA. **Top-left:** Likelihood scores across the entire dataset are tri-modal, with a significant proportion of responses between these modes; **Top-right:** likelihood scores for individual items typically approximate a Beta distribution with one mode; **Bottom-left:** items with median responses towards extreme ends of the scale are subject to higher inter-annotator agreement than for those in the middle ranges; **Bottom-right:** items with lower inter-annotator disagreement are (weakly) correlated with longer response times from participants.

datasets. Like ours, the human responses collected by Pavlick and Kwiatkowski (2019) for these datasets are on a numerical scale, and correspond to the likelihood of hypotheses in NLI items. Yet importantly, while they are also tri-modal, hardly any human responses lie between the three modes—indicating that the items in our dataset yield significantly more graded, probabilistic judgments than those in existing NLI datasets.

Human likelihood score distributions are almost always unimodal. While the overall distribution of likelihood scores across PROBCOPA is tri-modal, responses for *individual* items are almost always unimodal. Figure 2 (*top-right*) shows the distribution of human likelihood scores for one such item in our dataset. As it indicates, these responses approximate a Beta distribution with a single mode; we find that human responses for most items across the dataset do so as well. To confirm that our data is indeed unimodal, we use Silverman's (1981) statistical test of multimodality. The null hypothesis is that the sample distribution is unimodal; it is not rejected for any item in our dataset (at $\alpha = 0.05$).

The items in PROBCOPA therefore further stand out from those in existing NLI datasets, because while they are subject to judgment variation, the fact that this variation is unimodal—rather than multimodal—suggests that the items are not subject to qualitative differences in interpretation, as has been previously reported for NLI datasets (Pavlick and Kwiatkowski, 2019; Jiang et al., 2023).

Annotators do not collectively agree on a hypothesis having medium likelihood. Figure 2 (*bottom-left*) shows, for each item in our dataset, the differential entropy of human likelihood scores for the item (y-axis) plotted against its median likelihood score (x-axis). The differential entropy of human likelihood scores follows a horseshoe-like shape—items receiving high or low median likelihood scores are associated with comparatively lower differential entropy (i.e., higher inter-annotator agreement), while items with median likelihood scores closer to the middle of the scale are associated with higher differential entropy (i.e., lower inter-annotator agreement). Notably, we find no items for which annotators closely agree on a hypothesis having medium likelihood.

Higher entropy items are (weakly) correlated with longer human response times. Figure 2 (*bottom-right*) shows the differential entropy of

Downloaded from direct.mit.edu/tacl/article-pdf/doi/10.1162/TACL.a.791/2624521/tacl.a.791.pdf by guest on 20 September 2026

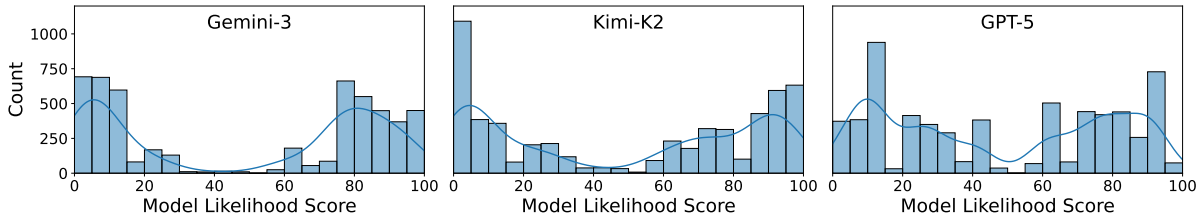


Figure 3: Distribution of likelihood scores across all PROBCOPA items, from three models. In contrast to humans (see Figure 2), models rarely return responses indicating medium likelihood, though this tendency is less extreme with GPT-5. See Figure 11 for the full set of distributions by model.

likelihood scores for each datapoint (y-axis), plotted against the mean time taken by participants to respond to that datapoint (log-transformed and z-scored; x-axis). We find a positive correlation between response time and likelihood score entropy (Spearman's $\rho = 0.31$, $p = 6.45e-06$)—meaning that, on average, items yielding lower inter-annotator agreement were also items that participants took longer to respond to in our experiment. We take this as evidence that the higher inter-annotator variation for some items is not a result of noise, but instead relates to item difficulty.

4 Comparison with Responses from Reasoning LLMs

Having analyzed how humans judge the probabilistic inferences in PROBCOPA, we now turn to LLMs. In particular, we test *reasoning LLMs*: LLMs that are trained to produce intermediate tokens (commonly referred to as a *reasoning chain*) before outputting a final response (Xu et al., 2025; Li et al., 2025; Marjanović et al., 2025).[3] We specifically focus on these models as (i) they represent the state-of-the-art on reasoning tasks, but (ii) are generally not evaluated on open-ended, non-deterministic reasoning contexts (see Section 7).

4.1 Methodology

Model Response Format We seek to obtain likelihood scores from reasoning models to compare with those we obtained from humans. While previous work has used model log-probabilities or sigmoid/softmax distributions in similar contexts (Pavlick and Kwiatkowski, 2019; Chen et al., 2020; Kauf et al., 2024), these are not accessible for most state-of-the-art reasoning LLMs. Moreover, since reasoning chains determine these models' final outputs, simply observing model probabilities conditioned on an input may fail to reflect actual output distributions when intermediate reasoning chains are generated. Conversely, while uncertainty quantification methods for black-box models may offer inspiration (e.g., Kuhn et al., 2023; Lin et al., 2024; Ulmer et al., 2024), these either (i) require gold labels against which accuracy can be measured, or (ii) are suited to open-ended generation settings. But probabilistic inferences by definition do not involve hard labels against which accuracy is a meaningful metric, and our setting involves likelihood estimates rather than open-ended generations.

For these reasons, following Mei et al. (2026), we obtain likelihood scores from reasoning LLMs via verbalized numerical estimates. For each item in our dataset, we ask the model to reason about the premise and hypothesis, and then return a value between 0 and 100 indicating the likelihood of the hypothesis given the premise. When doing so, we also provide the model with the same guide provided to humans describing how to partition the numerical scale (see Section 2.2). We repeat this 30 times for each item under the models' default temperature settings, to sample 30 likelihood scores per item for each model. The full prompt we provide to models is presented in Appendix C.

Metric for Distributional Comparison We again use differential entropy to quantify the spread of model likelihood scores (see Section 3.1). To make distributional comparisons between human and model scores, however, we need a metric of distributional similarity. We use *Wasserstein distance* (also known as *Earth Mover's Distance*). Formally, for two distributions P and Q, this is defined as:

$$(7) \quad W_1(P,Q) = \inf_{\gamma \in \Gamma(P,Q)} \mathbb{E}_{(x,y)\sim\gamma}[|x-y|]$$

[3] Some have criticized the use of terms like 'reasoning' and 'thinking' on grounds of anthropomorphization (Kambhampati et al., 2025). Note that while we follow convention in the field and use terms "reasoning LLMs", we do not suggest that reasoning chains are akin to human thoughts.

Downloaded from direct.mit.edu/tacl/article-pdf/doi/10.1162/TACL.a.791/2624521/tacl.a.791.pdf by guest on 20 September 2026

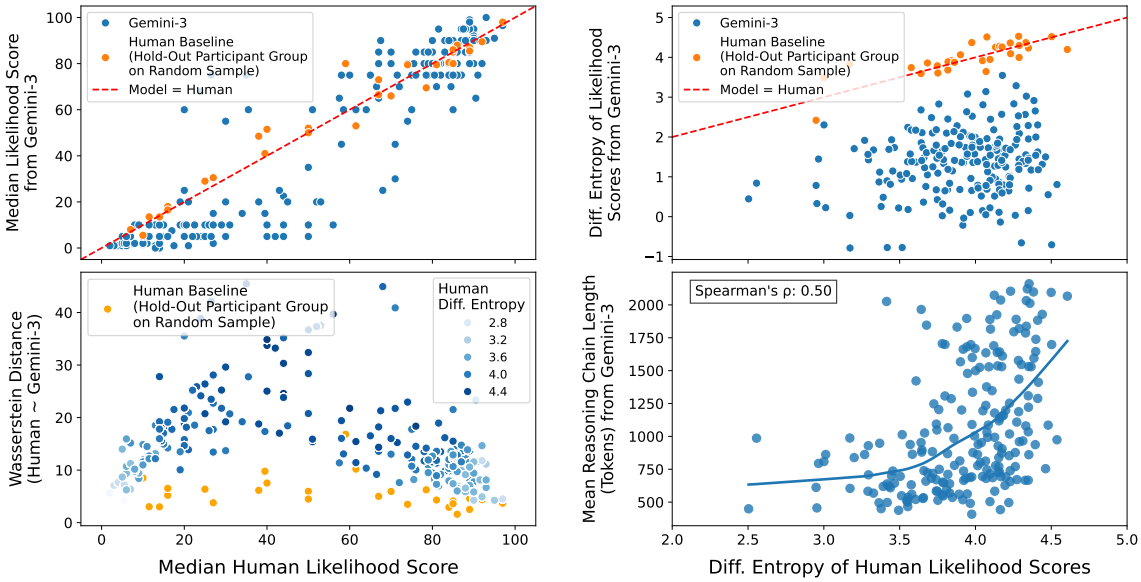


Figure 4: Item-wise comparisons between Gemini-3 and humans. **Top-left:** median likelihood scores from Gemini-3 align with those from humans at extreme ends of the scale, but not in the middle ranges; **Bottom-left:** likelihood score distributions from Gemini-3 and humans reflect the same pattern, with highest divergences for middle-range items (which also saw less inter-annotator agreement); **Top-right:** Gemini-3 shows less response diversity than humans for all items; **Bottom-right:** Gemini-3 on average reasons longer for items that humans disagree more on.

where $\Gamma(P, Q)$ denotes the set of all joint distributions with marginals P and Q. Intuitively, this captures the 'cost' of transforming one probability distribution into the other; higher values indicate lower distributional similarity, and lower values indicate higher similarity.[4]

Models Tested We test a range of contemporary reasoning LLMs from different providers: Gemini-3 (Gemini Team, 2025), GPT-5 (OpenAI, 2025a), Claude Sonnet-4.5 (Anthropic, 2025), Qwen3 (Qwen Team, 2025), Kimi-K2 (Kimi Team, 2025), GLM-4.6 (GLM-4.5 Team, 2025; Z.AI, 2025), DeepSeek-R1 (DeepSeekAI et al., 2025), and Grok-4.1 Fast (xAI, 2025). For exact model versions and inference details, see Appendix B.

Human Baseline When evaluating how closely model responses align with human likelihood scores, we also want a baseline of how well other humans can approximate these same scores. To establish this baseline, we therefore have a random sample of 30 PROBCOPA items re-annotated by a fresh set of participants, under the same annotation procedure reported in Section 2.2. When comparing a reasoning LLM's likelihood scores with those of PROBCOPA annotators, we then use this hold-out participant group's annotations to compute a baseline for human-to-human response similarity.

4.2 Results

Models rarely indicate medium likelihood. Figure 3 shows the overall distribution of likelihood scores (across all PROBCOPA items) for Gemini-3, Kimi-K2, and GPT-5. As it demonstrates, models exhibit a tendency not to return likelihood scores in the middle of the scale (i.e., those indicating medium likelihood). Though this tendency is least extreme for GPT-5 (see Figure 11 in Appendix E for the full set of model likelihood score distributions), it too rarely returns values in the very middle of the scale. Models thus appear committed to strong judgments of inference likelihood, supporting prior findings that they are often overconfident (see Section 7).

Model responses align with human responses more for low- and high-likelihood items than those in between. Figure 4 (*top-left*) shows the median human likelihood score (x-axis) against the median score from Gemini-3 (y-axis) for each

[4]We use this measure of distributional similarity over KL-divergence (another popular metric of distributional divergence), because unlike the latter, it does require the distributions to have matching support—model and human responses need not cover the same ranges of the likelihood scale.

Downloaded from direct.mit.edu/tacl/article-pdf/doi/10.1162/TACL.a.791/2624521/tacl.a.791.pdf by guest on 20 September 2026

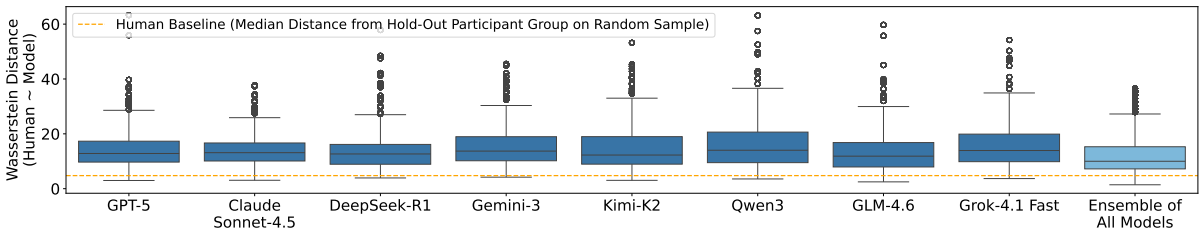


Figure 5: Distribution of item-wise Wasserstein distances between human and model likelihood score distributions. Ensembling the outputs of all models yields better distributional alignment with human judgments, but still falls short of the human-human baseline.

PROBCOPA item. As it suggests, while median responses from the model are similar to those from humans at the two extreme ends of the scale, this relationship breaks down closer to the middle of scale (since, as mentioned, models avoid responses in this range). As our baseline indicates, however, other humans are capable of reproducing similar median judgments for items across the scale.

We find this trend also holds when comparing entire distributions. Figure 4 (*bottom-left*) shows, for the same model, item-wise Wasserstein distances (y-axis) between human and model responses, as a function of the median likelihood score (x-axis) and differential entropy (color) from humans. As the plot suggests, distributional similarity between model and human likelihood scores is highest for items that humans collectively deem highly likely or unlikely, and lowest for items without such a consensus. Once again, however, we find no such pattern in our human baseline, which shows roughly the same degree of distributional similarity between our original and subsequent baseline annotations across all items. These trends hold for all models tested; full results are shown in Figures 12 and 13 (Appendix E).

Model responses almost never show as much variation as human responses. Figure 4 (*top-right*) compares, at an item-wise level, the differential entropy of likelihood scores from Gemini-3 (y-axis) and our original human annotators (x-axis). As it shows, for every single item in our dataset, likelihood scores from humans show higher differential entropy (in other words, more variation) than those from Gemini-3. Comparing with our human baseline, however, reveals roughly similar response variation between participants. Results for the full set of models tested are presented in Figure 14 in Appendix E; models almost never show more variation in their responses than humans.

We run follow-up experiments with the same random sample of 30 items we use for our human baseline, and find that adjusting temperature does not yield human-level response entropy from models; when increasing temperature, models will often devolve to generating endless sequences of random tokens before ever achieving human-like response variability. Similarly, we find that persona prompting (Zheng et al., 2024; Luz de Araujo et al., 2025) similarly has limited effects, and fails to deliver human-level response variation. These results are shown in Figures 15a and 15c, and support the general finding that state-of-the-art models often struggle to represent human variation (Santurkar et al., 2023; Zhang et al., 2026).

Increased reasoning effort does not significantly change models' likelihood scores. In the same follow-up study, we also assess whether increasing reasoning LLMs' *reasoning effort* parameters lead to significantly different outputs. For each item in human baseline random sample, we compare the median likelihood score produced under 'low' and 'high' reasoning effort parameters[5], and use bootstrapped 95% confidence intervals for these medians to check for statistical significance. Across all 30 items in the sample, and across all models, we never find any case of higher reasoning effort leading to a statistically significant difference in median likelihood score. Notably, these findings contrast those from Mei et al. (2026), who find that increased reasoning yields more overconfidence. See Figure 15b for further details.

Ensembling model responses leads to better (but not human-like) alignment with human response distributions. Figure 5 shows, for each

[5]For Gemini-3 and Claude Sonnet-4.5, we use thinking budgets of 512 and 4096 respectively to simulate the contrast between 'low' and 'high' reasoning effort.

Downloaded from direct.mit.edu/tacl/article-pdf/doi/10.1162/TACL.a.791/2624521/tacl.a.791.pdf by guest on 20 September 2026

Model	Premise	Hypothesis	Reasoning Chain Extract
Kimi-K2-Thinking	The tire on my bike was flat.	I pumped air into the tire.	...When a bike tire is flat, the most common and logical action is to pump air into it... However, there are other possibilities: ... They might not fix it at all... They might replace the tube/tire entirely... they might take it to a shop...
DeepSeek-R1	The girl looked pale.	Her father read her a story.	...This is an action a parent might take to comfort a child... However, it's not the only response. The father might giver her medicine, ask what's wrong, take her to a doctor, or just let her rest...
GLM-4.6	The man anticipated the team's victory.	He made a bet with his friends.	...when people anticipate a victory for a team they support, they might be more inclined to bet on that outcome... However... They might not be gamblers... They might not have money to bet... They might be satisfied with just watching the game...

Table 1: Sample excerpts of reasoning chains from different models, demonstrating the explicit considerations of alternative outcomes of the premise (highlighted in yellow).

model tested, the distribution of item-wise Wasserstein distances between model and human likelihood score distributions. As it demonstrates, model-human distributional differences are far higher than human-human distributional differences; ensembling model responses reduces this gap, but still falls short of the human baseline.

5 Analyzing LLM Reasoning Chains

Finally, we study reasoning LLMs' reasoning chains, to identify common patterns in how they reason over probabilistic inferences.

Models use more tokens on items subject to lower agreement among humans... Figure 4 (*bottom-right*) shows, for each PROBCOPA item, the mean reasoning chain length from Gemini-3 (y-axis; measured in tokens) against the differential entropy of human likelihood scores for that same item (x-axis). We see a clear correlation (Spearman's $\rho = 0.50$, $p = 2.10e-14$) between reasoning chain length and human differential entropy: on average, items that humans showed more uncertainty on yielded longer LLM reasoning chains. Table 4 in Appendix E shows these results for all models tested: most show at least modest correlations ($\rho \geq 0.30$), even if these are weaker than for Gemini-3.

...but correlations with human response time are much weaker. Despite this, we find that correlations between reasoning chain length and human response time (log-transformed and by-participant z-scored) are far lower. Table 4 shows these correlations for all models tested. While the highest correlation achieved between reasoning chain length and human differential entropy is 0.50 (from Gemini-3), the highest such correlation between reasoning chain length and human response time is only 0.25 (from Qwen-3). This indicates that while reasoning chain lengths may carry some relationship with how human judgments are distributed, relations to human cognitive load are far less clear.

Models explicitly reason over alternatives to arrive at likelihood judgments. For more qualitative insights into how models reason over PROBCOPA items, we manually inspect a random sample of 100 reasoning chains across all model responses. Doing so, we find a common pattern: 90 out of the 100 reasoning chains sampled include explicit considerations of alternative scenarios that are used to frame the model's final response. Table 1 shows some examples of this pattern, across different models. Though subject to questions of faithfulness (see Lanham et al., 2023; Xiong et al., 2026; Chen et al., 2025), the consistent use of alternative scenarios points to a common reasoning pattern across models, and invites further questions into how well this aligns with humans.

6 Discussion

Our results offer initial insights into how models reason in open-ended, non-deterministic settings, and point to the potential of further research in this area. For instance, our findings indicate that the tendency for models to be overconfident in their outputs (Mielke et al., 2022; Mei et al., 2026; Tian et al., 2023) is reflected even in open-ended inferences that are inherently uncertain; the models we test rarely indicate medium likelihood, and instead consistently favor more extreme likelihood scores. Similarly, our experiments reveal persistent differences between humans and models, with models failing to closely align with human judgment distributions, and producing far less variation in their responses than humans, even with different temperature settings (see Section 4.2). Such

Downloaded from direct.mit.edu/tacl/article-pdf/doi/10.1162/TACL.a.791/2624521/tacl.a.791.pdf by guest on 20 September 2026

issues of human-model similarity are of increasing importance as LLMs are used in human-focused settings (see e.g., Maity and Saikia, 2025; Wilcox et al., 2025; Anthis et al., 2025), and our work reiterates the need to assess models vis-à-vis these comparisons.

Conversely, our findings carry relevance for studies of human reasoning. Most notably, the graded, probabilistic judgments we see from our study participants (see Section 3.2) serve as empirical evidence of the probabilistic aspects of human reasoning and inference (Oaksford and Chater, 2007). Likewise, the observation that our human judgment distributions are unimodal stands out from recent work finding significant (often bimodal) human judgment variation towards NLI data (Pavlick and Kwiatkowski, 2019; Nie et al., 2020; Jiang et al., 2023), and suggests that sharp divergences in these judgments may not arise if the correct data is used (see also Jiang and de Marneffe, 2022).

More broadly, the data we collect has significant potential in the context of model calibration evaluation and probability estimation. For example, Jiang et al. (2024) argue that scalar annotations (such as ours) are a scalable way to estimate model calibration without sensitivity to different binning schemes (see Nixon et al., 2019). Similarly, Wang et al. (2025) present a framework for conditional probability estimation by fine-tuning models on probabilistic labels—and PROBCOPA offers high-quality human data for this exact setting.

7 Related Work

Reasoning in Humans Foundational work in modern mathematics and linguistics characterized inference patterns vis-à-vis formal logic (Frege et al., 1879; Tarski, 1936; Montague, 1970). Empirical research in psychology, however, has since suggested these are not in line with everyday human reasoning patterns, as several studies point to recurrent logical 'fallacies' from humans (e.g., Wason 1968; Evans et al. 1983, 1999; Klauer et al. 2000, see Evans 2002 for a review). Most notably, Wason (1968) found people frequently making faulty inferences from simple conditional statements, while Evans et al. (1983) showed that people often accept logically invalid arguments if their conclusions are believable. Oaksford and Chater (2007) thus argue that human reasoning should instead be understood in terms of probabilistic beliefs—a motivation we operationalize in this study.

Natural Language Inference In NLP, textual inferences are most often formalized via the *natural language inference* (NLI) task. Given a premise P and hypothesis H, the task traditionally involves classifying the pair as having an *entailment*, *contradiction* or *neutral* relation (Dagan et al., 2005).[6]

NLI has been used to study both specific types of inferences in NLP systems (e.g., Chen et al., 2020; Bhagavatula et al., 2020; Tian et al., 2021; Liu et al., 2023; Zhang et al., 2017; Jeretic et al., 2020) as well as general natural language understanding (see Poliak, 2020; Madaan et al., 2025). A growing body of research, however, reveals significant human judgment variation in NLI tasks (de Marneffe et al., 2012; Pavlick and Kwiatkowski, 2019; Nie et al., 2020; Jiang and de Marneffe, 2022; Jiang et al., 2023; Weber-Genzel et al., 2024; Kulmizev et al., 2026). Crucially, this line of work indicates that items in popular NLI datasets, such as SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018), are subject to judgment variation that goes beyond noise or crowdworker errors (Pavlick and Kwiatkowski, 2019; Weber-Genzel et al., 2024).

Closest to our work, Chen et al. (2020) study NLI probabilistically, re-annotating the SNLI dataset using a continuous scale. While their work is thus similar to ours in motivation, the data they use prevents the kind of analysis we conduct. As Nighojkar et al. (2023) note, the authors use the mean of only 2-3 crowdworker annotations as the gold label for each item. But SNLI items have been shown to yield bimodal judgment distributions (Pavlick and Kwiatkowski, 2019), making these averages potentially misleading; if one annotator judges an inference as highly likely, and the other judges the opposite, the resulting mean would indicate medium likelihood, despite no annotator believing this. These limitations motivate us to construct our own dataset, detailed in Section 2.

Reasoning LLMs Reasoning LLMs, like OpenAI's o3 (OpenAI, 2025b) or DeepSeek AI's DeepSeek-R1 (DeepSeekAI et al., 2025), are trained to produce intermediate tokens (known as a *reasoning chain* or *thinking trace*) before outputting a final response (Xu et al., 2025; Li et al., 2025; Marjanović et al., 2025). These LLMs present strong reasoning capabilities, with major gains on code and reasoning benchmarks (OpenAI,

[6]Entailment and contradiction in NLI often do not align with strict logical entailment and contradiction; see Zaenen et al. (2005); Manning (2006); Crouch et al. (2006) for more.

Downloaded from direct.mit.edu/tacl/article-pdf/doi/10.1162/TACL.a.791/2624521/tacl.a.791.pdf by guest on 20 September 2026

2024; DeepSeekAI et al., 2025; Kimi Team et al., 2025; Qwen Team et al., 2025; Liu et al., 2025a).

Crucially, however, this work is largely centered around mathematical and logical reasoning. The training pipelines behind reasoning LLMs typically involve math or coding tasks that are automatically verifiable (Lambert et al., 2024; Liu et al., 2025a). Similarly, in evaluation, reasoning capabilities are typically assessed via math and coding datasets like AIME (Mathematical Association of America, 2024) and SWE-BENCH (Jimenez et al., 2024) (see e.g., DeepSeekAI et al., 2025; OpenAI, 2024, 2025b; Kimi Team et al., 2025; Qwen Team et al., 2025).

Some work *has* looked at how LLMs reason with probabilities (Marzoev et al., 2026; Pournemat et al., 2025; Paruchuri et al., 2024; Xia et al., 2024; Nafar et al., 2025). But this work mostly focuses on probability theory and explicit statistical distributions (e.g., calculating percentiles), rather than open-ended reasoning contexts that may not involve explicit probability theory (see e.g., Table 1). A notable exception is from Wang et al. (2025), who focus on LLMs' conditional probability estimation abilities in a setting closer to ours, finding that fine-tuning LLMs on probabilistic labels improves these abilities in models. Our work is complementary to these efforts, as we present high-quality, human-annotated data that can act as reliable ground-truth labels for this task.

Uncertainty Quantification for LLMs Uncertainty quantification (UQ) in the context of LLMs asks how certain or confident models are of their outputs, typically in contrast to some measure of how certain or confident they *should* be (Ulmer, 2024; Liu et al., 2025b; Shorinwa et al., 2025).

Since we are only interested in the likelihood LLMs assign to probabilistic inferences—rather than their uncertainty around such estimates—our work is slightly outside the scope of traditional UQ methods (see Lin et al., 2024). Nevertheless, some work in the domain is highly relevant to our analysis.

Several studies have examined how LLMs explicitly verbalize uncertainty, both through numerical estimates or linguistic markers (Lin et al., 2022; Yona et al., 2024; Tian et al., 2023; Belém et al., 2024, see Ulmer et al. 2025 for an overview). Much of this line of work finds that LLMs are *overconfident*, often being more confident of their outputs than is warranted (Mielke et al., 2022; Tian et al., 2023; Krause et al., 2023; Mei et al., 2026). Closest to our study, Mei et al. (2026) find that reasoning LLMs are typically overconfident, and that deeper reasoning from models leads to greater overconfidence. Interestingly, while our results align with LLMs being overconfident, they do not show any clear effects of reasoning effort (see Section 4.2).

8 Conclusion

In this paper, we assessed probabilistic reasoning in both humans and LLMs, using PROBCOPA, a novel dataset of 210 probabilistic inferences in English, each with at least 25 human annotations. We find significant differences between how humans and reasoning LLMs judge probabilistic inferences, with models failing to match human judgment distributions or produce human-level output variation. Furthermore, we analyze model reasoning chains, and identify common reasoning patterns, but mixed correlations with human behavior. Our findings raise further questions about *why* these trends arise across reasoning LLMs, and invite future studies into the causal effects of different stages of pre- and post-training on this downstream behavior. More generally, we hope our work inspires further research on reasoning in open-ended, human-like and non-deterministic contexts.

Limitations

Besides being limited to English, our study is subject to other limitations we highlight below.

Verbalized Likelihood Scores While we argue that reasoning LLMs are best-suited to verbalized likelihood scores for the purposes of our study (see Section 4.1), questions remain around how faithful these generally are (Tian et al., 2023; Kumar et al., 2024). We thus hope that future work identifies other methods for likelihood elicitation that are suited to the specific nature of reasoning models.

COPA-Derived Items Our dataset is novel, but its items are derived from COPA (Roemmele et al., 2011), an older dataset that likely features in the training data of most models. Since our re-framing of the task around these items yields new judgments compared to the original COPA gold labels (see Section 2.1), we do not believe that we are testing on a task the model has already been trained on; nevertheless, it is possible that the presence of the some of these sentences in the training data affects model behavior towards them.

Downloaded from direct.mit.edu/tacl/article-pdf/doi/10.1162/TACL.a.791/2624521/tacl.a.791.pdf by guest on 20 September 2026

Acknowledgments

The authors would like to thank Verna Dankers, Marius Mosbach, Dennis Ulmer, Ivan Titov and Desmond Elliot for providing crucial feedback on this work, as well as the TACL action editor and reviewers. This work was also made possible with the support of the IVADO *R3 NLP Régroupement*, the Canada CIFAR AI Chair and the NSERC Discovery Grant. Gaurav Kamath is supported by a Doctoral Training Award from the *Fonds de Recherche du Québec – Société et Culture*. Marie-Catherine de Marneffe is a Research Associate of the *Fonds de la Recherche Scientifique – FNRS*. Sebastian Schuster has been supported by the Vienna Science and Technology Fund (WWTF) [10.47379/VRG23007] *Understanding Language in Context.* Every word in this paper was written by a human.

References

Jacy Reese Anthis, Ryan Liu, Sean M Richardson, Austin C Kozlowski, Bernard Koch, Erik Brynjolfsson, James Evans, and Michael S Bernstein. 2025. Position: LLM social simulations are a promising research method. In *Forty-second International Conference on Machine Learning Position Paper Track*.

Anthropic. 2025. Claude Sonnet 4.5 system card. Technical report, Anthropic. Accessed: 2025-12-30.

Catarina Belém, Markelle Kelly, Mark Steyvers, Sameer Singh, and Padhraic Smyth. 2024. Perceptions of linguistic uncertainty by language models and humans. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8467–8502.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*.

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

On Climate Change et al. 2007. Intergovernmental panel on climate change. *World Meteorological Organization*, 52(1-43):1.

Tongfei Chen, Zheng Ping Jiang, Adam Poliak, Keisuke Sakaguchi, and Benjamin Van Durme. 2020. Uncertain natural language inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8772–8779.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vlad Mikulik, Samuel R. Bowman, Jan Leike, Jared Kaplan, and Ethan Perez. 2025. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*.

Richard Crouch, Lauri Karttunen, and Annie Zaenen. 2006. Circumscribing is not excluding: A response to Manning. Unpublished manuscript.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL recognising textual entailment challenge. In *Machine learning challenges workshop*, pages 177–190. Springer.

Marie-Catherine de Marneffe, Christopher D Manning, and Christopher Potts. 2012. Did it happen? The pragmatic complexity of veridicality assessment. *Computational Linguistics*, 38(2):301–333.

DeepSeekAI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

David M. Eddy. 1982. Probabilistic reasoning in clinical medicine: Problems and opportunities. *Judgment under uncertainty: Heuristics and biases*, pages 249–267.

Jonathan St BT Evans. 2002. Logic and human reasoning: an assessment of the deduction paradigm. *Psychological bulletin*, 128(6):978.

Jonathan St BT Evans, Julie L Barston, and Paul Pollard. 1983. On the conflict between logic

Downloaded from direct.mit.edu/tacl/article-pdf/doi/10.1162/TACL.a.791/2624521/tacl.a.791.pdf by guest on 20 September 2026

and belief in syllogistic reasoning. *Memory & cognition*, 11(3):295–306.

Jonathan St BT Evans, Simon J Handley, Catherine NJ Harper, and Phillip N Johnson-Laird. 1999. Reasoning about necessity and possibility: A test of the mental model theory of deduction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25(6):1495.

Gottlob Frege et al. 1879. Begriffsschrift, a formula language, modeled upon that of arithmetic, for pure thought. *From Frege to Gödel: A source book in mathematical logic*, 1931:1–82.

Gemini Team. 2025. Gemini 3 Pro model card. Technical report, Google DeepMind. Accessed: 2025-12-30.

GLM-4.5 Team. 2025. GLM-4.5: Agentic, reasoning, and coding (ARC) foundation models. *arXiv preprint arXiv:2508.06471*.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705, Online. Association for Computational Linguistics.

Nan-Jiang Jiang and Marie-Catherine de Marneffe. 2022. Investigating reasons for disagreement in natural language inference. *Transactions of the Association for Computational Linguistics*, 10:1357–1374.

Nan-Jiang Jiang, Chenhao Tan, and Marie-Catherine de Marneffe. 2023. Ecologically valid explanations for label variation in NLI. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10622–10633.

Zhengping Jiang, Anqi Liu, and Benjamnin Van Durme. 2024. Addressing the binning problem in calibration assessment through scalar annotations. *Transactions of the Association for Computational Linguistics*, 12:120–136.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. 2024. SWE-bench: Can language models resolve real-world GitHub issues? In *The Twelfth International Conference on Learning Representations*.

Subbarao Kambhampati, Kaya Stechly, Karthik Valmeekam, Lucas Paul Saldyt, Siddhant Bhambri, Vardhan Palod, Atharva Gundawar, Soumya Rani Samineni, Durgesh Kalwar, and Upasana Biswas. 2025. Stop anthropomorphizing intermediate tokens as reasoning/thinking traces! In *NeurIPS 2025 Workshop on Bridging Language, Agent, and World Models for Reasoning and Planning*.

Carina Kauf, Emmanuele Chersoni, Alessandro Lenci, Evelina Fedorenko, and Anna Ivanova. 2024. Log probabilities are a reliable estimate of semantic plausibility in base and instruction-tuned language models. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 263–277.

Kimi Team. 2025. Kimi K2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. 2025. Kimi k1.5: Scaling reinforcement learning with LLMs. *arXiv preprint arXiv:2501.12599*.

Karl Christoph Klauer, Jochen Musch, and Birgit Naumer. 2000. On belief bias in syllogistic reasoning. *Psychological review*, 107(4):852.

Lea Krause, Wondimagegnhue Tufa, Selene Báez Santamaría, Angel Daza, Urja Khurana, and Piek Vossen. 2023. Confidently wrong: exploring the calibration and expression of (un) certainty of large language models in a multilingual setting. In *Proceedings of the workshop on multimodal, multilingual natural language generation and multilingual WebNLG Challenge (MM-NLG 2023)*, pages 1–9.

Tevye R Krynski and Joshua B Tenenbaum. 2007. The role of causality in judgment under uncertainty. *Journal of Experimental Psychology: General*, 136(3):430.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*.

Downloaded from direct.mit.edu/tacl/article-pdf/doi/10.1162/TACL.a.791/2624521/tacl.a.791.pdf by guest on 20 September 2026

Artur Kulmizev, Erika Lombart, Patrick Watrin, and Marie-Catherine de Marneffe. 2026. Label and explanation variation in LLM-based annotation: a case study in natural language inference. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16526–16543, San Diego, California, United States. Association for Computational Linguistics.

Abhishek Kumar, Robert Morabito, Sanzhar Umbet, Jad Kabbara, and Ali Emami. 2024. Confidence under the hood: An investigation into the confidence-probability alignment in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 315–334.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. 2024. Tülu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. 2023. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*.

Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, et al. 2025. From system 1 to system 2: A survey of reasoning large language models. *arXiv preprint arXiv:2502.17419*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*.

Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah A Smith, and Yejin Choi. 2023. We're afraid language models aren't modeling ambiguity. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 790–807.

Keliang Liu, Dingkang Yang, Ziyun Qian, Weijie Yin, Yuchi Wang, Hongsheng Li, Jun Liu, Peng Zhai, Yang Liu, and Lihua Zhang. 2025a. Reinforcement learning meets large language models: A survey of advancements and applications across the LLM lifecycle. *arXiv preprint arXiv:2509.16679*.

Xiaoou Liu, Tiejin Chen, Longchao Da, Chacha Chen, Zhen Lin, and Hua Wei. 2025b. Uncertainty quantification and confidence calibration in large language models: A survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 6107–6117.

Pedro Henrique Luz de Araujo, Paul Röttger, Dirk Hovy, and Benjamin Roth. 2025. Principled personas: Defining and measuring the intended effects of persona prompting on task performance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26857–26886, Suzhou, China. Association for Computational Linguistics.

Lovish Madaan, David Esiobu, Pontus Stenetorp, Barbara Plank, and Dieuwke Hupkes. 2025. Lost in inference: Rediscovering the role of natural language inference for large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9229–9242.

Downloaded from direct.mit.edu/tacl/article-pdf/doi/10.1162/TACL.a.791/2624521/tacl.a.791.pdf by guest on 20 September 2026

Kyle Mahowald, Peter Graff, Jeremy Hartman, and Edward Gibson. 2016. SNAP judgments: A small N acceptability paradigm (SNAP) for linguistic acceptability judgments. *Language*, 92(3):619–635.

Subhankar Maity and Manob Jyoti Saikia. 2025. Large language models in healthcare and medical applications: A review. *Bioengineering*, 12(6):631.

Christopher D. Manning. 2006. The PASCAL RTE1 challenge. Unpublished manuscript.

Sara Vera Marjanović, Arkil Patel, Vaibhav Adlakha, Milad Aghajohari, Parishad BehnamGhader, Mehar Bhatia, Aditi Khandelwal, Austin Kraft, Benno Krojer, Xing Han Lù, et al. 2025. DeepSeek-R1 thoughtology: Let's think about LLM reasoning. *arXiv preprint arXiv:2504.07128*.

Alana Marzoev, Jacob Andreas, and Jillian Ross. 2026. Openestimate: Evaluating llms on reasoning under uncertainty with real-world data. In *The Fourteenth International Conference on Learning Representations*.

Mathematical Association of America. 2024. AIME 2024: American invitational mathematics examination. https://maa.org/math-competitions/aime. Accessed November 2025.

Zhiting Mei, Christina Zhang, Tenny Yin, Justin Lidard, Ola Sho, and Anirudha Majumdar. 2026. Reasoning about uncertainty: Do reasoning models know when they don't know? In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 3408–3458, Rabat, Morocco. Association for Computational Linguistics.

Sabrina J Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872.

Richard Montague. 1970. Universal grammar. *Theoria*, 36(3).

Aliakbar Nafar, Kristen Brent Venable, Zijun Cui, and Parisa Kordjamshidi. 2025. Extracting probabilistic knowledge from large language models for Bayesian network parameterization. *arXiv preprint arXiv:2505.15918*.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143, Online. Association for Computational Linguistics.

Animesh Nighojkar, Antonio Laverghetta Jr, and John Licato. 2023. No strong feelings one way or another: Re-operationalizing neutrality in natural language inference. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 199–210.

Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. 2019. Measuring calibration in deep learning. In *CVPR Workshops*, volume 2.

Mike Oaksford and Nick Chater. 2007. *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press.

OpenAI. 2024. Learning to reason with LLMs. https://openai.com/index/learning-to-reason-with-llms/. Accessed: 2025-11-03.

OpenAI. 2025a. GPT-5 system card. Technical report, OpenAI. Accessed: 2025-12-30.

OpenAI. 2025b. OpenAI o3 and OpenAI o4-mini system card. https://openai.com/index/o3-o4-mini-system-card/. System card.

Akshay Paruchuri, Jake Garrison, Shun Liao, John Hernandez, Jacob Sunshine, Tim Althoff, Xin Liu, and Daniel McDuff. 2024. What are the odds? Language models are capable of probabilistic reasoning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11712–11733.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Adam Poliak. 2020. A survey on recognizing textual entailment as an NLP evaluation. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 92–109, Online. Association for Computational Linguistics.

Downloaded from direct.mit.edu/tacl/article-pdf/doi/10.1162/TACL.a.791/2624521/tacl.a.791.pdf by guest on 20 September 2026

Mobina Pournemat, Keivan Rezaei, Gaurang Sriramanan, Arman Zarei, Jiaxiang Fu, Yang Wang, Hamid Eghbalzadeh, and Soheil Feizi. 2025. Reasoning under uncertainty: Exploring probabilistic reasoning capabilities of LLMs. *arXiv preprint arXiv:2509.10739*.

Qwen Team. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Qwen Team, An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, pages 90–95.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR.

Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z Ren, and Anirudha Majumdar. 2025. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *ACM Computing Surveys*.

Bernard W Silverman. 1981. Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society: Series B (Methodological)*, 43(1):97–99.

Jon Sprouse, Carson T Schütze, and Diogo Almeida. 2013. A comparison of informal and formal acceptability judgments using a random sample from Linguistic Inquiry 2001–2010. *Lingua*, 134:219–248.

Jean-Louis Stilgenbauer, Jean Baratgin, and Igor Douven. 2017. Reasoning strategies for diagnostic probability estimates in causal contexts: Preference for defeasible deduction over abduction. In *DARe LPNMR*.

Alfred Tarski. 1936. Der Wahrheitsbegriff in den formalisierten Sprachen. *Studia philosophica*, 1.

Jidong Tian, Yitian Li, Wenqing Chen, Liqiang Xiao, Hao He, and Yaohui Jin. 2021. Diagnosing the first-order logical reasoning ability through LogicNLI. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3738–3747.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442.

Dennis Ulmer, Martin Gubri, Hwaran Lee, Sangdoo Yun, and Seong Oh. 2024. Calibrating large language models using their generations only. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15440–15459.

Dennis Ulmer, Alexandra Lorson, Ivan Titov, and Christian Hardmeier. 2025. Anthropomimetic uncertainty: What verbalized uncertainty in language models is missing. *arXiv preprint arXiv:2507.10587*.

Dennis Thomas Ulmer. 2024. *On Uncertainty in Natural Language Processing*. Phd thesis, IT University of Copenhagen, Copenhagen, Denmark.

Gaëlle Villejoubert and David R Mandel. 2002. The inverse fallacy: An account of deviations from Bayes's theorem and the additivity principle. *Memory & cognition*, 30(2):171–178.

Liaoyaqi Wang, Zhengping Jiang, Anqi Liu, and Benjamin Van Durme. 2025. Always tell me the odds: Fine-grained conditional probability estimation. In *Second Conference on Language Modeling*.

Peter C Wason. 1968. Reasoning about a rule. *Quarterly journal of experimental psychology*, 20(3):273–281.

Leon Weber-Genzel, Siyao Peng, Marie-Catherine de Marneffe, and Barbara Plank. 2024. VariErr NLI: Separating annotation error from human label variation. In *Proceedings of the 62nd*

Downloaded from direct.mit.edu/tacl/article-pdf/doi/10.1162/TACL.a.791/2624521/tacl.a.791.pdf by guest on 20 September 2026

Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2256–2269.

Ethan Gotlieb Wilcox, Michael Y Hu, Aaron Mueller, Alex Warstadt, Leshem Choshen, Chengxu Zhuang, Adina Williams, Ryan Cotterell, and Tal Linzen. 2025. Bigger is not always better: The importance of human-scale language modeling for psycholinguistics. *Journal of Memory and Language*, 144:104650.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122.

Bonnie C Wintle, Hannah Fraser, Ben C Wills, Ann E Nicholson, and Fiona Fidler. 2019. Verbal probabilities: Very likely to be somewhat more confusing than numbers. *PLoS One*, 14(4):e0213522.

xAI. 2025. Grok 4.1 model card. Technical report, xAI. Accessed: 2025-12-30.

Shepard Xia, Brian Lu, and Jason Eisner. 2024. Let's think Var-by-Var: Large language models enable ad hoc probabilistic reasoning. *arXiv preprint arXiv:2412.02081*.

Zidi Xiong, Shan Chen, Zhenting Qi, and Himabindu Lakkaraju. 2026. Measuring the faithfulness of thinking drafts in large reasoning models. *Advances in Neural Information Processing Systems*, 38:139438–139467.

Fengli Xu, Qianyue Hao, Chenyang Shao, Zefang Zong, Yu Li, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, Yuwei Yan, Qinglong Yang, Yiwen Song, Sijian Ren, Xinyuan Hu, Jie Feng, Chen Gao, and Yong Li. 2025. Toward large reasoning models: A survey of reinforced reasoning with large language models. *Patterns*, 6(10):101370.

Gal Yona, Roee Aharoni, and Mor Geva. 2024. Can large language models faithfully express their intrinsic uncertainty in words? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7752–7764.

Annie Zaenen, Lauri Karttunen, and Richard Crouch. 2005. Local textual inference: can it be defined or circumscribed? In *Proceedings of the ACL workshop on empirical modeling of semantic equivalence and entailment*, pages 31–36.

Z.AI. 2025. GLM-4.6. https://z.ai/blog/glm-4.6. Blog post. Accessed: 2025-12-30.

Lily H Zhang, Smitha Milli, Karen Long Jusko, Jonathan Smith, Brandon Amos, Wassim Bouaziz, Manon Revel, Jack Kussman, Yasha Sheynin, Lisa Titus, Bhaktipriya Radharapu, Jane Yu, Vidya Sarma, Kristopher Rose, and Maximilian Nickel. 2026. Cultivating pluralism in algorithmic monoculture: The Community Alignment dataset. In *The Fourteenth International Conference on Learning Representations*.

Sheng Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2017. Ordinal commonsense inference. *Transactions of the Association for Computational Linguistics*, 5:379–395.

Mingqian Zheng, Jiaxin Pei, Lajanugen Logeswaran, Moontae Lee, and David Jurgens. 2024. When "a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15126–15154, Miami, Florida, USA. Association for Computational Linguistics.

Downloaded from direct.mit.edu/tacl/article-pdf/doi/10.1162/TACL.a.791/2624521/tacl.a.791.pdf by guest on 20 September 2026

A PROBCOPA Human Annotation Procedure

Human annotators recruited via Prolific participated in our crowdsourced experiment that we ran using a custom website built on HTML and JavaScript. After reviewing and accepting a consent form, participants were presented with instructional examples that demonstrated the task format. Figure 6 shows the first such instructional example: participants are presented with the general task format, and requested to provide a response using the slider. Note that we framed hypotheses as *possible effects* of premises due to the original setting of COPA items, as well as because doing so offers an intuitive interpretation of inference likelihood. Upon submitting a response, they would receive automatic feedback based on the range in which they responded. In this phase, we aimed to use simple examples for which most people would share a broad consensus on likelihood ranges. Figure 7 shows an example of such feedback, if participants provided the 'wrong' response (participants would also receive positive feedback if they responded within the intended ranges of the scale). Participants were presented with 5 such instructional examples that familiarized them with low, middle, and high ranges of the scale.

Upon completion of this instructional phase, participants were informed that they would now enter the main phase of the experiment, for which there were no 'right' or 'wrong' answers. Figure 8 shows an example of the UI in this main phase. Participants were presented with up to 30 PROBCOPA items sequentially (with similarly formatted attention checks interspersed); they provided responses for each, and were given no further feedback. At the end of the experiment, participants were given the chance to raise any comments or questions about how the experiment was conducted; we received no feedback indicating any difficulty with the task.

As mentioned in Section 2.3, we also ran two rounds of human validation after obtaining our original annotations. In the first, we re-ran the exact same experiment with 30 new participants (on a subset of the data); in the second, we also adjusted the prompt wording slightly. Figure 9 shows an example from this second round of human validation, where we align the exact task wording more closely with the prompt provided to LLMs (see Appendix C). Note that in both rounds of human validation, we obtained response distributions that were not statistically significantly different from our original annotations (see Section 2.3).

B Model Inference Details

Model Name	Exact Model Version
Gemini-3	`gemini-3-pro-preview`
GPT-5	`gpt-5-2025-08-07`
Claude Sonnet-4.5	`claude-sonnet-4-5-20250929`
Qwen3	`Qwen3-235B-A22B-Thinking-2507`
Kimi-K2	`Kimi-K2-Thinking`
GLM-4.6	`GLM-4.6`
Grok-4.1 Fast	`grok-4.1-fast`
DeepSeek-R1	`DeepSeek-R1`

Table 2: Exact model versions used in this study.

Table 2 shows the exact models used in this study. We ran inference on Gemini-3 using the Gemini API; GPT-5 using the OpenAI API; Claude Sonnet-4.5 using the Anthropic API; and Qwen3, Kimi-K2, GLM-4.6 and DeepSeek-R1 using the Together AI API. For all of these models, we made API calls using each respective provider's Batch API functionality. We ran inference on Grok-4.1 Fast using OpenRouter's API (which did not offer a Batch API functionality).

Temperature values were set to model defaults except when running temperature experiments (see Figure 15a, Section 4.2). Reasoning effort was set to 'medium' for GPT-5, Qwen3, Kimi-K2, GLM-4.6 and DeepSeek-R1 (which take this argument) while 'thinking budget' was set to 1024 for Claude Sonnet-4.5 and Gemini-3 (which take this argument instead)—once again, except when running reasoning effort experiments (see Section 4.2). Note that although the Gemini API documentation suggests that the model accepts a 'reasoning effort' parameter, at the time of running our experiments, this had not been implemented in the batch API functionality—which is why we instead controlled the 'thinking budget' parameter for the model. Maximum new token limits were set to 2048 for the main experiment and persona prompting experiments, and increased to 4224 for the temperature and reasoning effort experiments.

Downloaded from direct.mit.edu/tacl/article-pdf/doi/10.1162/TACL.a.791/2624521/tacl.a.791.pdf by guest on 20 September 2026

C Prompt to Models

Below is the exact prompt we provide to the reasoning LLMs we test in this study. Note that the scale provided in the system prompt is the same scale provided to humans in the annotation procedure.

Prompt:

```
Consider the following situation and
possible effect.

Situation: {premise}
Possible Effect: {hypothesis}

Given the situation, how likely is
this effect? Respond with a numerical
value between 0 and 100, where 0
indicates that this is DEFINITELY NOT
the effect, and 100 indicates that this
is DEFINITELY the effect.
```

System Prompt:

```
You provide responses to questions
about the likelihood of an effect given
some situation. After any internal
reasoning, reply with a single number
between 0 and 100, enclosed in <answer>
tags. You can use the following
descriptions of numerical ranges to
help guide your response:
0: Absolutely no chance
1-5: Almost no chance
6-15: Highly unlikely
16-34: Unlikely
35-49: Somewhat unlikely
50: Totally even chance
51-65: Somewhat likely
66-84: Likely
85-94: Highly likely
95-99: Almost certain
100: Absolutely certain
```

D Persona Prompting

As mentioned in Section 4.2, we run a follow-up experiment on a subset of PROBCOPA, in which we prompt the models with different persona descriptions, to test whether this yields more human-like response distributions. For each of the 30 responses we sample from a model on a single PROBCOPA item (see Section 3.1), we append to the system prompt (see Appendix C) a different persona description, that specifies either a demographic or psychological description. See Table 3 for examples. As Figure 15c shows, such persona prompting fails to provide human-level response variation or human-like response distributions.

Persona Prompt Examples
Demographic: You are a 23-year-old female barista in Ottawa, who is saving up to backpack across Europe. Your first language is English.
Demographic: You are a 58-year-old male factory worker in Detroit, who is nearing retirement after three decades in the auto industry. Your first language is English.
Psychological: You balance curiosity with pragmatism, comfortable with both the familiar and the new. You are dependable and organized, often making plans to reach your goals. You feel comfortable in both quiet and social settings, adapting as needed. You are caring and cooperative, often attuned to the feelings of those around you. You experience typical ups and downs but manage emotions with balance.

Table 3: Examples of persona descriptions used in our persona prompting experiments. Neither approach yielded human-level variation or human-like response distributions (see Figure 15c).

E Extended Results Figures

Figures 10 to 14 and 15a to 15c below show extended results referred to in the main body of this paper.

Downloaded from direct.mit.edu/tacl/article-pdf/doi/10.1162/TACL.a.791/2624521/tacl.a.791.pdf by guest on 20 September 2026

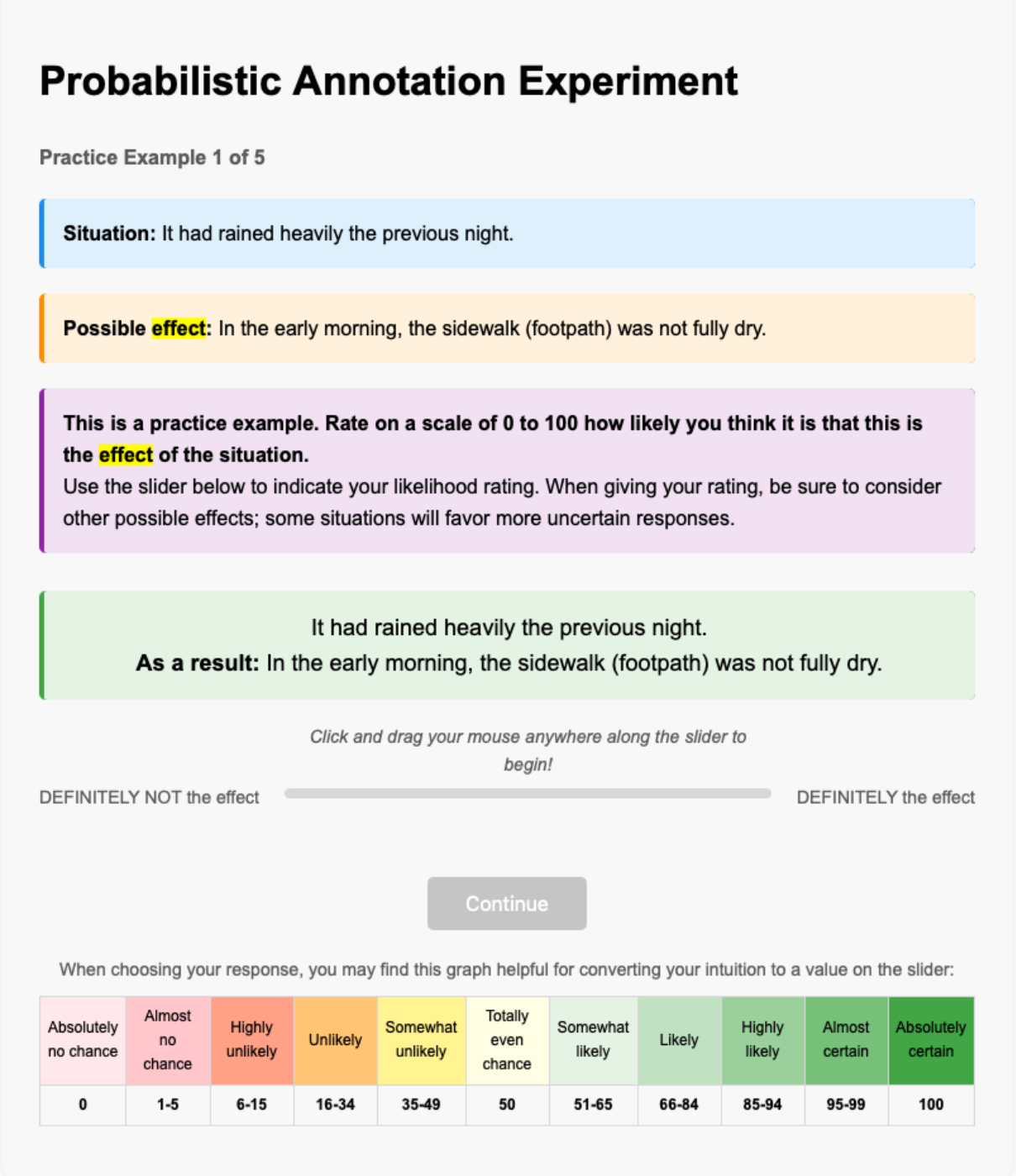


Figure 6: Screenshot of the first instructional example presented to participants in our crowdsourced experiment. Participants were shown the general task format, and asked to present a response using the slider. The guide presented beneath the example was intended to align participants on how to use the scale. In this instructional stage, participants were given automatic feedback based on their responses.

Downloaded from direct.mit.edu/tacl/article-pdf/doi/10.1162/TACL.a.791/2624521/tacl.a.791.pdf by guest on 20 September 2026

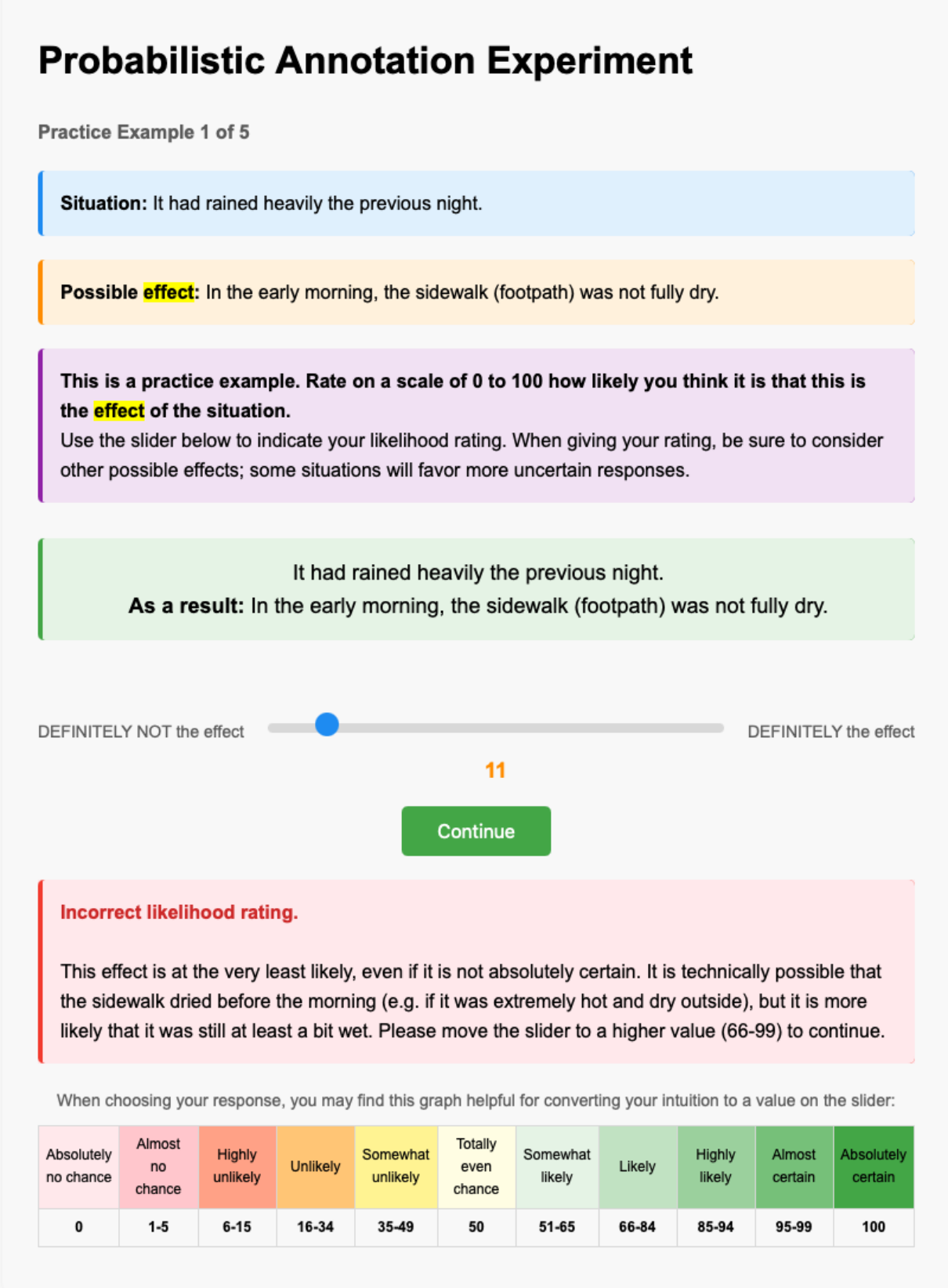


Figure 7: Screenshot showing the automatic feedback participants would receive if they provided a likelihood score outside of the 'likely' to 'almost certain' range for the first instructional example (see Figure 6). In this stage, we aimed to use simple examples for which most people would agree on broad likelihood ranges.

Downloaded from direct.mit.edu/tacl/article-pdf/doi/10.1162/TACL.a.791/2624521/tacl.a.791.pdf by guest on 20 September 2026

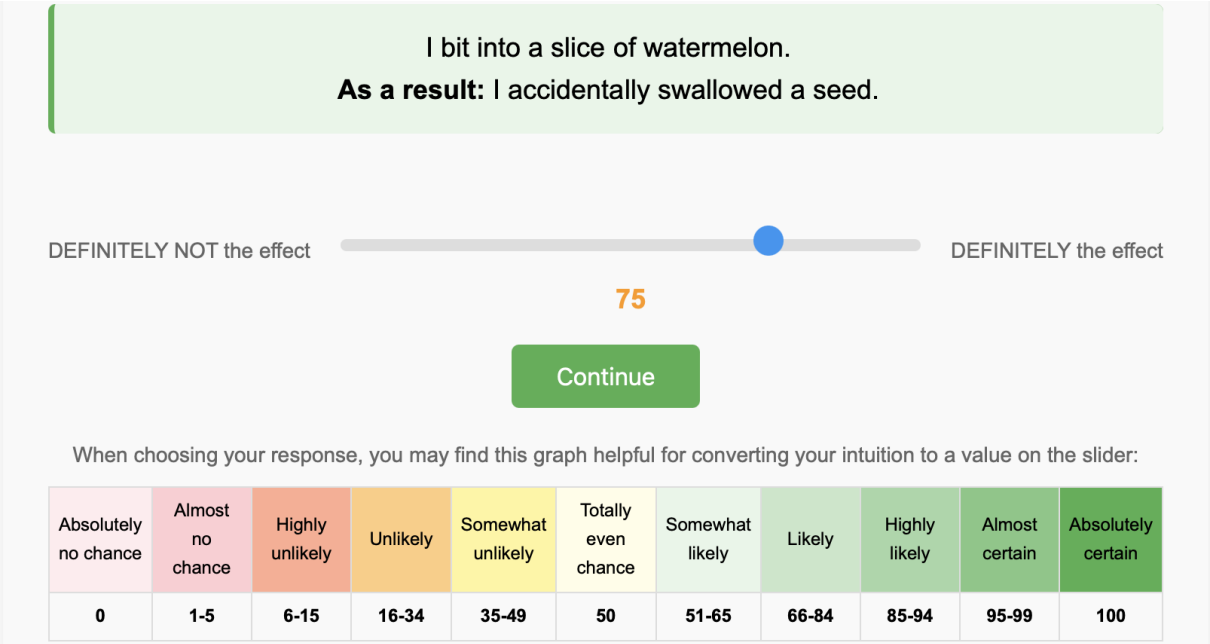


Figure 8: Screenshot of the annotation UI for the main phase of the crowdsourced experiment. The UI and task format follow from what participants were shown in the instructional phase. But at this stage, participants have been informed that unlike in the previous phase, there are no 'right' or 'wrong' answers.

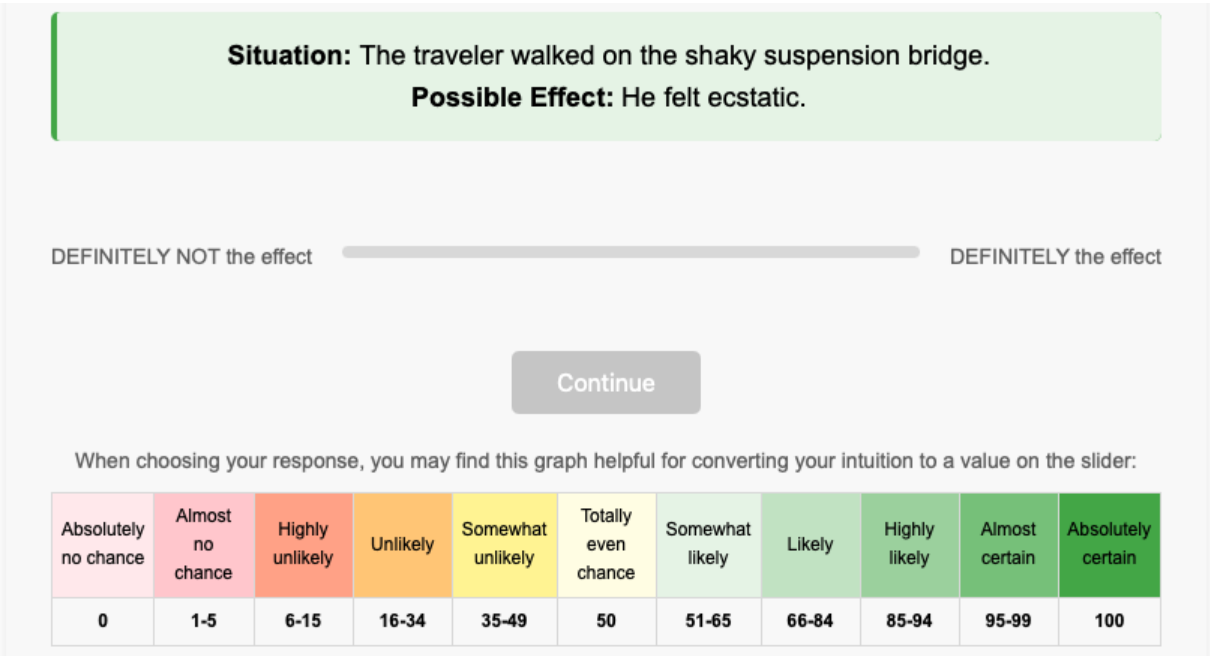


Figure 9: Screenshot of the annotation UI for our validation experiment, in which we slightly vary the prompt wording to closer align with the wording models are presented with (see Section 2.3). Note that this variation does not produce different response distributions from our original annotations.

Downloaded from direct.mit.edu/tacl/article-pdf/doi/10.1162/TACL.a.791/2624521/tacl.a.791.pdf by guest on 20 September 2026

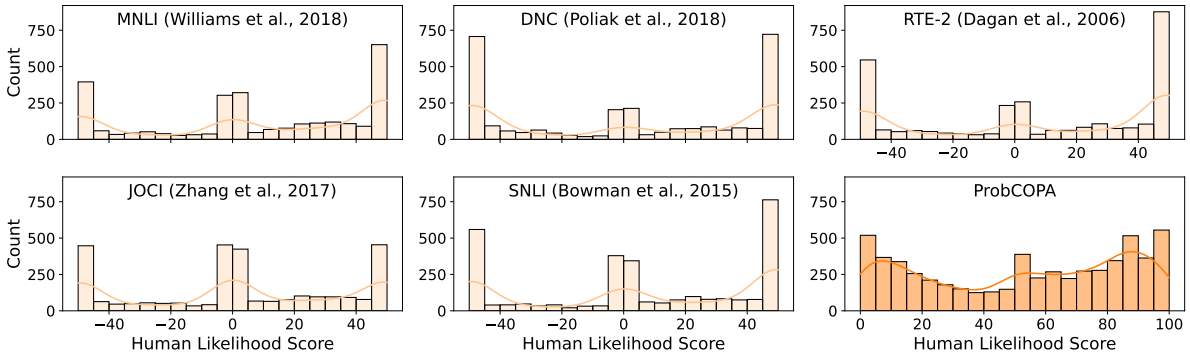


Figure 10: Overall human likelihood score distribution across (i) five major NLI datasets, collected by Pavlick and Kwiatkowski (2019), and (ii) PROBCOPA (ours). Likelihood scores collected by Pavlick and Kwiatkowski (2019) for the five major NLI datasets lie on a scale from -50 (hypothesis definitely false given premise) to 50 (hypothesis definitely true given premise). All datasets are subject to tri-modal distributions; but PROBCOPA items receive far more annotations that lie in between these three modes, indicating more graded, probabilistic judgments than for other NLI datasets.

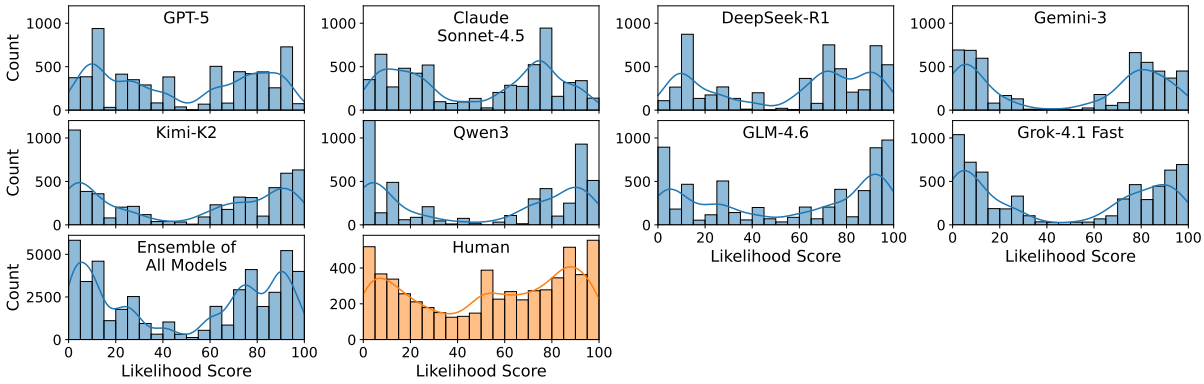


Figure 11: Distribution of likelihood scores across all PROBCOPA items, from all models tested, contrasted against the same distribution from humans. While humans yield an overall likelihood score distribution that is tri-modal, with a large number of responses towards the middle range of the scale, models yield an overall distribution that is bi-modal, with few likelihood scores in the middle range of the scale.

	Reasoning chain length (tokens) $\sim$ Differential entropy of human likelihood scores		Reasoning chain length (tokens) $\sim$ Human response time (log-transformed and z-scored)	
Model	Spearman's ρ	p-value	Spearman's ρ	p-value
GPT-5	0.30	7.06e-06	0.17	1.19e-02
Claude Sonnet-4.5	0.18	9.61e-03	0.12	8.74e-02
DeepSeek-R1	0.36	6.54e-08	0.18	9.85e-03
Gemini-3	0.50	2.10e-14	0.18	9.86e-03
Kimi-K2	0.33	1.05e-06	0.24	4.44e-04
Qwen3	0.27	6.97e-05	0.25	2.48e-04
GLM-4.6	0.14	4.18e-02	-0.02	8.22e-01
Grok-4.1 Fast*	NA	NA	NA	NA
Ensemble of All Models	0.44	1.90e-11	0.23	6.35e-04

Table 4: Spearman correlations between reasoning chain lengths and (i) the differential entropy of human likelihood scores, and (ii) human response time, log-transformed and by-participant z-scored. While correlations between reasoning chain lengths and human likelihood score entropy suggest a relationship between the two for most models, correlations with human response time are consistently lower. *Grok-4.1 Fast does not return reasoning chain information, and is therefore excluded from this analysis.

Downloaded from direct.mit.edu/tacl/article-pdf/doi/10.1162/TACL.a.791/2624521/tacl.a.791.pdf by guest on 20 September 2026

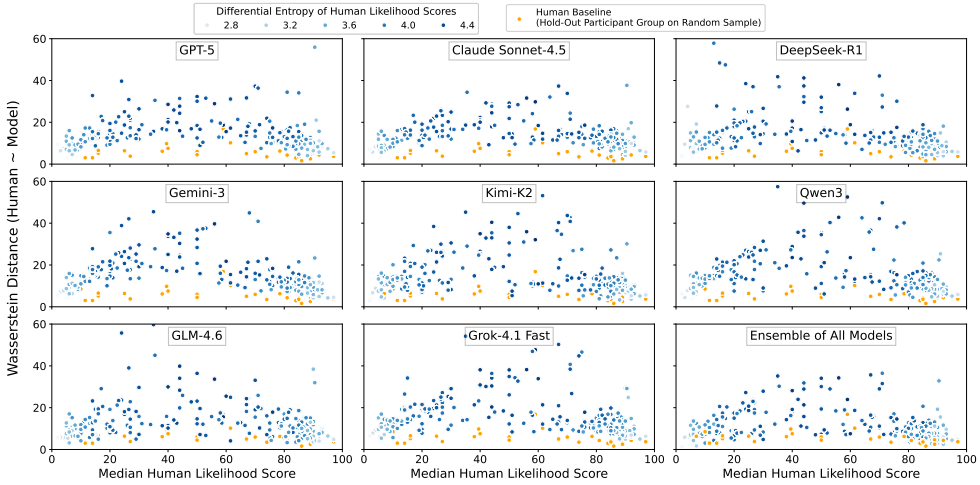


Figure 12: Item-wise model-human Wasserstein distances, for all models tested. Unlike the human-human baseline, models increasingly diverge from human distributions on middle-range items with high entropy.

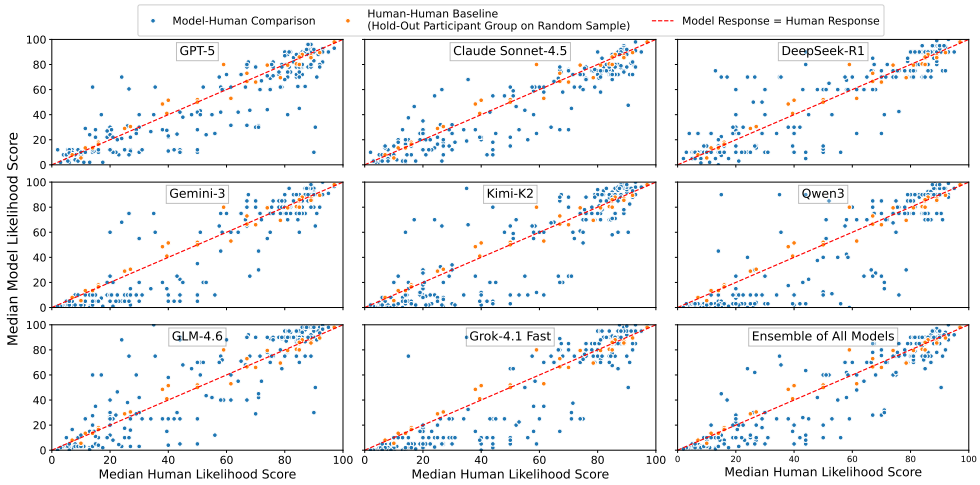


Figure 13: Item-wise model and human median likelihood scores, for all models tested. Median ratings are similar at the two ends of the scale, but break down towards the middle.

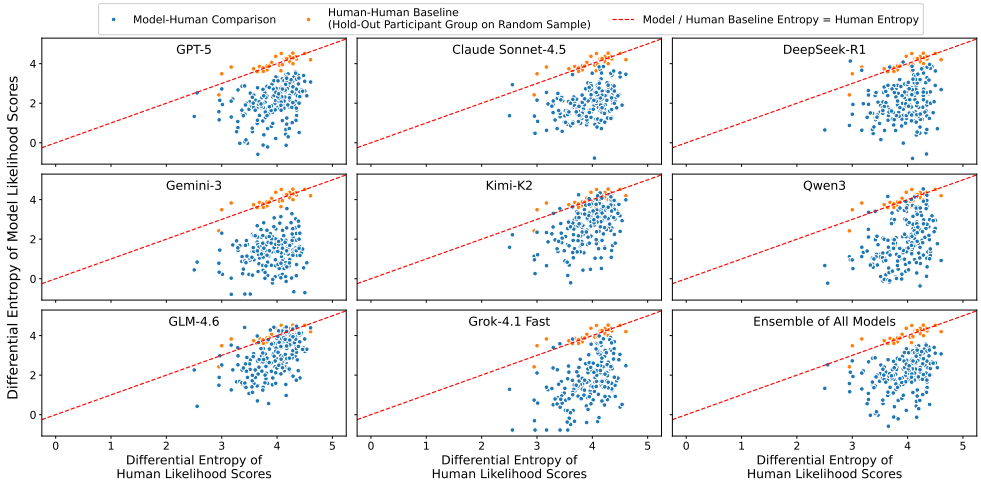
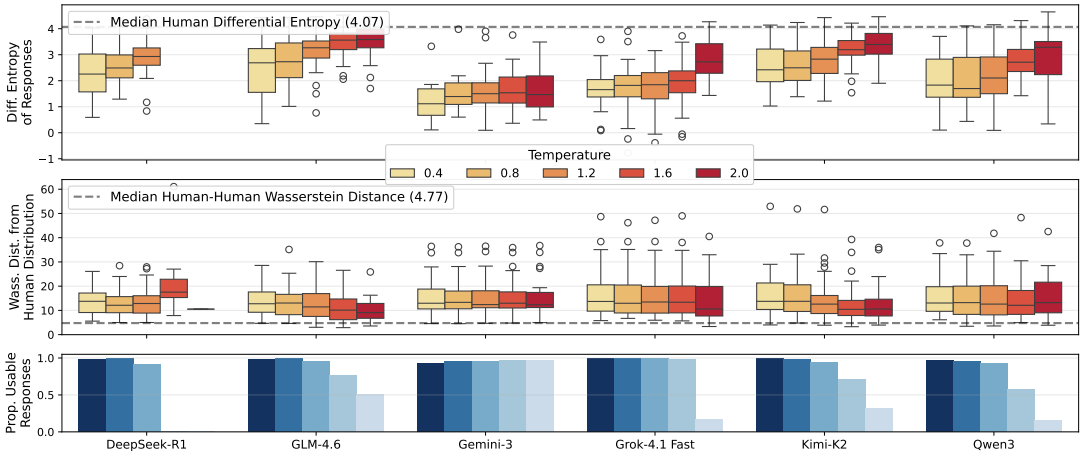
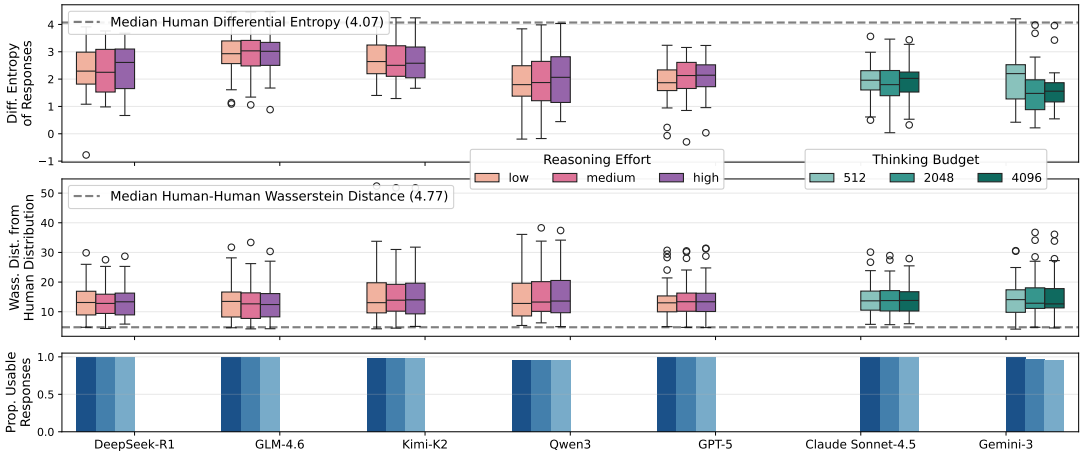


Figure 14: Item-wise differential entropy of model and human likelihood scores, for all models tested. Human response distributions almost universally show higher entropy than model response distributions.

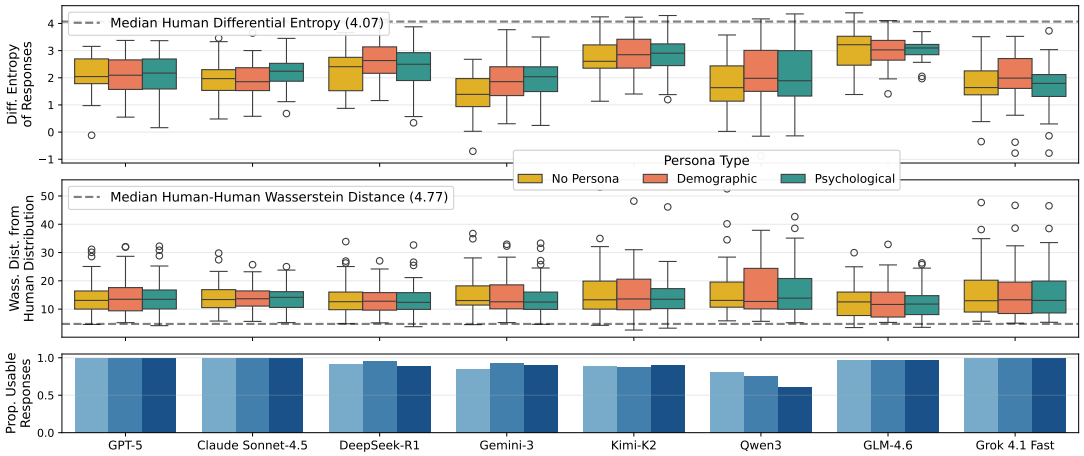
Downloaded from direct.mit.edu/tacl/article-pdf/doi/10.1162/TACL.a.791/2624521/tacl.a.791.pdf by guest on 20 September 2026



(a) **Temperature**



(b) **Reasoning Effort**



(c) **Persona Prompting**

Figure 15: Results showing the effect of adjusting model temperature (15a), reasoning effort (15b), and persona prompting (15c). Each panel shows, under different conditions, the distribution of item-wise differential entropy values (**top rows**), Wasserstein distances from human responses (**middle rows**), and proportion of parseable responses returned before hitting new token limits (**bottom rows**). Increasing temperature yields slightly more diverse responses, but at the cost of output quality. Changes to reasoning effort or persona prompting fail to yield more human-like distributions or variation.

Downloaded from direct.mit.edu/tacl/article-pdf/doi/10.1162/TACL.a.791/2624521/tacl.a.791.pdf by guest on 20 September 2026