Perceptions of Linguistic Uncertainty by Language Models and Humans

Anonymous ACL submission

Abstract

Uncertainty expressions such as "probably" or "highly unlikely" are pervasive in human language. While prior work has established that there is general population-level agreement among humans about what these expressions mean quantitatively, the abilities of LLMs to interpret these phrases have seen little investigation. In this paper, we introduce a task for evaluating the abilities of LLMs to interpret uncertainty expressions as probabilities. Our approach assesses whether LLMs can employ theory of mind in this setting: understanding the uncertainty of another agent about a particular statement, independently of the LLM's own certainty about that statement. We evaluate both humans and a variety of LLMs on this task, demonstrating that a variety of LLMs are able to map uncertainty expressions to probabilistic responses in a human-like manner. However, we observe systematically different behavior depending on whether a statement is actually true or false. This sensitivity indicates that LLMs are substantially more susceptible to bias based on their prior knowledge (as compared to humans). These findings raise crucial questions and have broad implications for human-AI and AI-AI communication of uncertainty.

1 Introduction

Uncertainty is ubiquitous in human communication — in relaying predictions ("it is likely to rain tomorrow"), conveying imperfect knowledge ("I think I have a copy in my desk"), and describing unknown information ("the artifact could be more than 500 years old"). Expressing uncertainty is critical in fields such as medicine, law, and politics, where statements including *uncertainty expressions* (e.g., "likely," "doubtful") are frequently used to support medical, judicial, and political decisions (Karelitz and Budescu, 2004). For instance, domain experts use these expressions to provide imprecise likelihood assessments about the side-effects of a medical treatment (Sawant and Sansgiry, 2018; Patt and Dessai, 2005), the chances of winning a not-guilty verdict in legal cases (Fore, 2019), the probability of environmental events resulting from climate change (Patt and Dessai, 2005; Ho et al., 2015), or the likelihood of emergence of military conflicts (Duke, 2023). Generally, humans are well-attuned to such statements, exhibiting population-level agreement in mapping these expressions to corresponding probabilities (Wallsten et al., 1986a; Willems et al., 2019; Fagen-Ulmschneider, 2019).

However, the ability of large language models (LLMs) to understand linguistic uncertainty has received relatively little attention. In particular, given text where a speaker expresses uncertainty about a particular statement, we are interested in whether LLMs can interpret the uncertainty not as a function of their internal beliefs, but by objectively assessing the speaker's uncertainty about the statement. Consider the motivating example in Figure 1: when writing a headline for a statement qualified by the word "probable," ChatGPT expresses substantially different uncertainty depending on its prior belief about the statement.[1] In this example, ChatGPT is conflating the speaker's uncertainty with its own uncertainty about the statement—in effect, a failure of "theory of mind."

In this work, we investigate the abilities of LLMs to provide quantitative interpretations of uncertainty expressions, focusing in particular on how the prior knowledge of an LLM affects this ability. To this end, we introduce a new task[2] in which LLMs must map text containing uncertainty expressions to numerical probabilities. We analyze the performance of both humans and several popular LLMs on this task, enabling direct comparison between humans and models. We find that larger,

[1]ChatGPT agrees with the first statement and disagrees with the second; see Figure 9 in the Appendix.

[2]A link to our dataset and code will be made available upon acceptance.

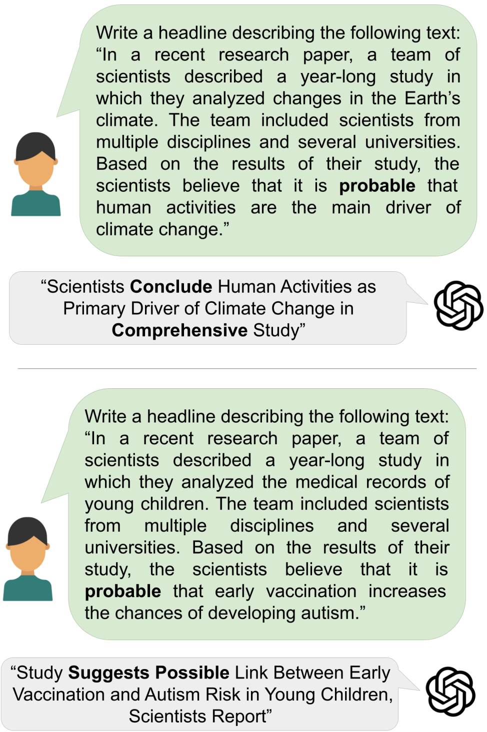


Figure 1: Two interactions with ChatGPT (June 2024). In each, ChatGPT is asked to write a headline for a short passage. Both passages are structured identically and qualified with the word "probable," but the first is about climate change and the second about the link between vaccines and autism. For the first passage, ChatGPT generates a certain-sounding headline, using words like "conclude" and "comprehensive." The second headline is weaker, with words like "suggests" and "possible."

newer models can consistently map uncertainty expressions to numerical probabilities that align with human-like perceptions. However, we also show that the probabilities LLMs choose are susceptible to bias based on their prior knowledge—to a much greater extent than those of humans.

This propensity has concerning implications given the increasing use of LLMs for generating content (e.g., summarization, data augmentation) and evaluating language generation (Wang et al., 2023). When an LLM's ability to quantify uncertainty can be "poisoned" by its beliefs, its downstream performance is dependent on its parametric or pretraining knowledge (which can be obsolete or wrong (Liang et al., 2022; Longpre et al., 2023)), rather than on critical contextual information (Longpre et al., 2021). Further, this means that the biases of a model (including the many well-documented potentially harmful biases of LLMs, e.g., Wan et al. (2023); Kotek et al. (2023); Salewski et al. (2024); Scherrer et al. (2024); Motoki et al. (2024)) can subtly manifest in how it interprets and generates uncertainty language.

2 Related Work

Human Perceptions of Uncertainty Expressions. In fields like medicine, finance, law, and politics, where it is impossible to make predictions with complete certainty, decisions are often informed by subjective probabilities (Karelitz and Budescu, 2004; Dhami and Wallsten, 2005; Fore, 2019). Subjective probabilities can be communicated quantitatively, through numerical probabilities (e.g., odds, percentages, intervals), or qualitatively, through the use of uncertainty expressions or epistemological markers (e.g., "I believe", "According to") (Dhami and Mandel, 2022). Although being less precise than numerical probabilities (Wallsten et al., 1986b; Brun and Teigen, 1988; Budescu et al., 2014), humans generally prefer to use linguistic expressions, rather than numbers, to communicate uncertainty (Erev and Cohen, 1990; Wallsten et al., 1993).

Interested in the efficacy of how humans communicate uncertainty linguistically, researchers have examined how participants map uncertainty expressions into numerical values across different fields and expertise levels (Karelitz and Budescu (2004); Wallsten et al. (2008, 1986a); Fore (2019); *inter alia*). Although there can be considerable variation in responses at the individual level, these studies have revealed that there are consistent patterns relating uncertainty expressions and numerical probabilities that can be observed systematically at the population level (Wallsten et al., 2008; Willems et al., 2019; Fagen-Ulmschneider, 2019).

Uncertainty Quantification in LLMs. The need for more reliable LLMs has prompted researchers to investigate new methods for communicating internal uncertainty of LLMs. Proposed methods can be differentiated in terms of the information used to estimate the model confidence in its response. For instance, some methods leverage information about the token-level logits of generated outputs (Jiang et al., 2021; Kuhn et al., 2023; Duan et al., 2024), resort to sampling multiple responses (Si et al., 2022; Chen and Mueller, 2023; Xiong et al., 2024; Hou et al., 2024; Lin et al., 2024; Aichberger et al., 2024), train external classifiers to produce confidence estimates based on the inputs and/or LLMs' representations (Jiang et al., 2021; Mielke et al., 2022; Shrivastava et al., 2023), or elicit these confidences directly from LLMs as output tokens (Lin et al., 2022; Tian et al., 2023). While these works investigate how LLMs express uncertainty when generating text, there has been far less work on the question we focus on in this paper, i.e., how LLMs interpret uncertainty in text.

More recently, concerns about human over-reliance on LLMs spurred investigations about the impact of LLM-articulated uncertainty in human-AI interaction (Zhou et al., 2024; Kim et al., 2024; Steyvers et al., 2024). After conducting human studies, the authors find that participants tend to rely less on LLMs when their outputs include uncertainty expressions. By assessing LLMs' perceptions of linguistic uncertainty, our work aims to understand the impact that uncertainty expressions have in model behavior.

Most directly related to our work is that of Maloney et al. (2024), who compare numerical probability estimates from GPT-4 and humans using a small set of "context" prompts. Our paper goes significantly beyond this work by assessing a broad range of LLMs using a more diverse and natural set of contexts. Further, we evaluate in a "theory of mind" context, prompting LLMs to estimate what an uncertainty expression reflects about the speaker's belief, rather than what the expression means to the LLM. In addition, our work is the first that we are aware of to investigate how LLMs can be biased by their prior knowledge in mapping uncertainty expressions to numerical probabilities.

3 Baseline Human Study

As a baseline for how people map uncertainty expressions to numerical probabilities, we first conduct an experiment in which humans are shown uncertainty expressions and are asked to provide corresponding numerical probability estimates. We focus on a set of 14 uncertainty expressions (e.g., "almost certain", "unlikely"—see the full list in Figure 4), drawn from Wallsten et al. (2008) and Wallsten et al. (1986a). In this initial experiment, our goal is to capture how people perceive these phrases "in the wild," putting them in the context of real-world statements. An additional goal is to select statements that minimize the potential for people to conflate their own beliefs about these statements with their assessment of the confidence of the person making the statement. To this end, we construct a set of statements (u, s, e) which include uncertainty expressions $u \in \mathcal{U}$ used by speakers $s \in \mathcal{S}$ to convey their degree of certainty about the truthfulness or falsehood of a statement or event $e \in \mathcal{E}$. This degree of certainty can be expressed by a number between 0 and 100, where 0 implies that a speaker s believes there is a 0% chance that e is true whilst 100 implies a belief that there is a 100% chance that the statement is true.

By presenting statements about a specific speaker s (with a random name), we are asking participants to use "theory of mind," estimating the belief of someone else. We can then query participants about the speaker's strength of belief, clearly distinguishing this notion from the participant's own beliefs. For instance, given the statement "Sonia believes it is unlikely it will rain today," we can ask participants to quantify how likely *Sonia* thinks it is to rain, distinct from the probability the participant themselves believes it will rain.

We use *non-verifiable* statements e to separate the meaning of the uncertainty expressions from uncertainty about the statements themselves. These are statements that are not sufficiently grounded with specific contextual information to allow an external observer to be confident in either the truth or falsity of the statement. For example, in the context of a prompt such as "Maria believes it is likely that [statement]", statements such as *her boss has two pets* or *her flight will land around 6pm* are statements we consider non-verifiable in the sense that there is insufficient context provided (that the speaker knows, but does not state) for an observer to be able to reliably assess the likelihood that the statement is true. In contrast, *verifiable* statements (which we discuss further in Section 4.1) can be verified as correct or incorrect in a context-free sense (e.g., *the capital city of Peru is Lima*); humans and LLMs will often have strong prior beliefs about the likelihood that such statements are true.

For this baseline experiment, we manually constructed a set of 60 non-verifiable statements and systematically combined these with the aforementioned 14 uncertainty phrases. We randomly generated speaker names, generating sentences describing the belief of a hypothetical speaker in the form: "[Speaker] believes it is [uncertainty phrase] that [statement]." For each of these sentences, participants are asked to quantify the speaker's belief about the statement, in particular, they were asked what is the probability being expressed *from the speaker's perspective* that the statement is correct.

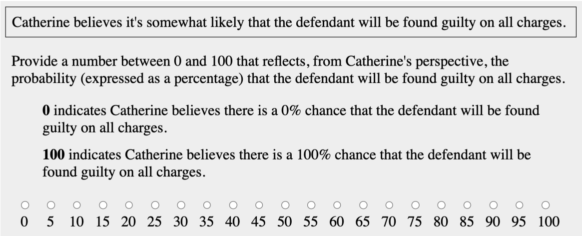


Figure 2: Example question provided to participants in the baseline experiment.

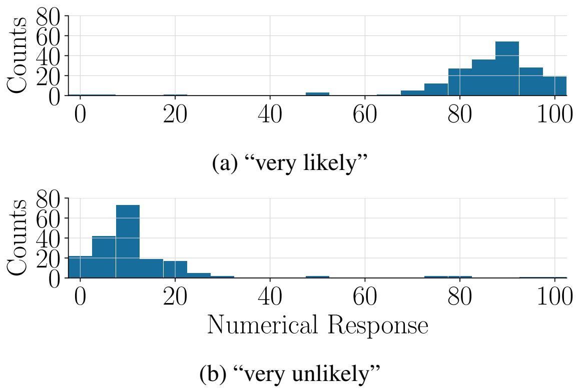


Figure 3: Histogram of participant responses for non-verifiable statements for two uncertainty expressions.

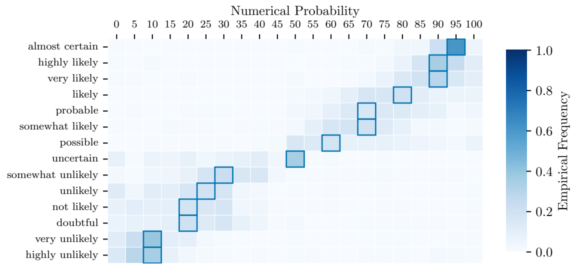


Figure 4: Human empirical distributions of numerical responses per uncertainty expression in the non-verifiable setting. Highlighted blue boxes represent the maximum value for each expression.

Participants then provided their response quantized to numerical bins $0, 5, 10, \ldots, 95, 100$. An example of what was shown to participants in the experiment is shown in Figure 2. Each of the 94 participants in this experiment generated responses in this manner for two randomly selected statements (and speaker names) for each of the 14 different uncertainty expressions.

The result of this experiment[3] is a distribution over the probabilities participants associate with each uncertainty expression. For example, Figure 3 reflects the probabilities assigned to the phrases "very likely" and "very unlikely"; results for all 14 uncertainty expressions are shown in Figure 4. Overall, we observe similar results to prior work on these perceptions (Wallsten et al., 2008, 1986a; Willems et al., 2019), including consistent ordering in aggregate population patterns, as determined by the mode of the empirical distribution.

4 Methodology

4.1 Verifiable Statements

In addition to the non-verifiable statements described in Section 3, our dataset also includes *verifiable* statements, for the purpose of assessing the effects of prior knowledge on quantifying linguistic uncertainty. We curate 60 verifiable statements based on a multiple-choice question-answering dataset from The Question Company.[4] Starting with 30 of the dataset's "easy" questions and corresponding multiple-choice options, we write *true* statements that use the correct answer and *false* statements using one of the incorrect answers. In our main paper, we focus on results using these 60 statements.[5] Examples of statements and details about the dataset are included in Appendix C.

4.2 Numerical responses from LLMs

To obtain uncertainty estimates from LLMs, we create prompts similar to the queries provided to humans (see Appendix C). For a given statement (u, s, e), we need to estimate a distribution over the LLM's generated numerical probabilities. In this paper we consider two techniques for obtaining this distribution: greedy sampling and (when available) the next-token probability distribution.

In greedy sampling, we approximate the empirical distribution with a single sampled response (`temperature=0`)—an approach commonly used to solve discriminative tasks with LLMs (Zhu and Griffiths, 2024). Because this sampling approach requires no knowledge about the weights or next-token probabilities, it is applicable to any model, including those behind black-box APIs (e.g., `Gemini` (Anil et al., 2024), `GPT-4` (Achiam et al., 2024)). We focus primarily on this greedy sampling approach since it aligns more closely with the human responses, i.e., we want to assess the ability of each

[3] Additional details about our human experiments can be found in Appendix A.

[4] `https://www.thequestionco.com/`

[5] For further validation, our dataset includes 400 additional statements (generated from the AI2-ARC question set (Clark et al., 2018) via a similar procedure)—we report results on this full set in Appendix E.

LLM as if it were an individual human providing responses rather than asking it to match a population distribution of human responses.

We obtain an empirical distribution of probabilities conditional on an uncertainty expression u by we repeating this process over multiple statements e and speakers s.

4.3 Metrics

We treat the empirical distribution obtained for the non-verifiable statements (see Section 3) as the *reference distribution*, as it reflects inter-human variability in a setting designed to be free of prior information or biases about the corresponding statements. For every uncertainty expression $u \in \mathcal{U}$, we define a reference conditional probability distribution $P(k|u)$, over values of k that are multiples of 5 in the range $[0, 100]$, where $P(k|u)$ is as the empirical distribution from the baseline experiment with non-verfiable statements. Given a response from any agent, human or LLM, we measure the quality of the agreement of the response with the reference distribution.

The primary quality metric that we use is **Proportional Agreement (PA)**, defined as follows. If an agent selects bin k for uncertainty expression u then the PA value for that response is defined as $P(k|u)$, where P is the reference (population) distribution. Intuitively, for an expression u, this PA score $P(k|u)$ represents the probability that the agent's response k agrees with that of a randomly selected individual, and is upper bounded for any expression by $\arg\max_k P(k|u)$, i.e., by the mode of the $P(k|u)$. The higher the PA, the better the quality of the response in terms of agreement with the aggregate human population (as reflected by $P(k|u)$. To get a single score for an LLM or individual human, we average PA over multiple responses and over the 14 uncertainty expressions.

Note that the PA metric is similar to the log-probability metric widely used to score probabilistic models in machine learning. However it is not a likelihood in the sense of a model assigning probability mass to observed data—in this context it is appropriate to average the PA scores directly (rather than taking products of probabilities as would be done under an IID likelihood assumption). An alternative to the PA metric would be to compare histograms of responses, e.g., based on multiple responses from agents for a particular uncertainty expression u. We provide numerical results (using the Wasserstein distance between histograms) in the Appendix, but this is of secondary interest since we are not requiring any LLM or individual human to necessarily replicate the full population variability of responses.

As an additional measure of alignment between the reference distribution and the agent's distribution, we also compute the **Mean Absolute Error (MAE)** over the means of the empirical distributions for each uncertainty expression, i.e. we compute the absolute difference between the means for each expression, and then average across the expressions.

5 Results

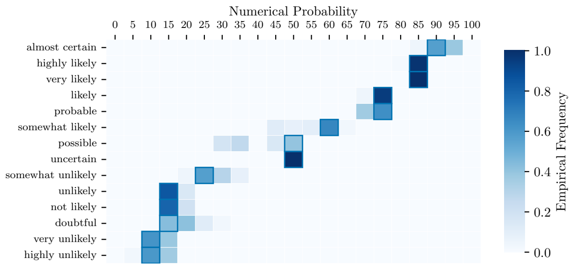
This section examines the ability of several well-known LLMs to interpret uncertainty expressions.[6] We begin by assessing models' abilities to produce numerical responses that resemble human-like trends (e.g., higher numerical responses assigned to higher certainty expressions and vice-versa). We then study the effect of prior knowledge in the perception of uncertainty of both humans and models. We conclude with some results on the generalizability of our findings.

5.1 How well do LLMs perceive uncertainty?

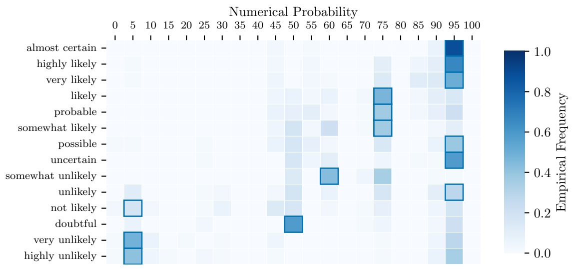
As established in prior work and in our baseline experiment, humans show population-level agreement in mapping uncertainty expressions to numerical probabilities. In this section, we assess whether LLMs possess a similar ability to ascribe numerical probabilities to uncertainty expressions. To this end, we prompt LLMs to provide numerical probabilities for the same non-verifiable (NV) statements as in the baseline experiment (Section 3). In Figure 5 we include expression-wise histograms for these numerical probabilities for `GPT-4o` and `OLMo (7B)` (which can be compared to the histogram for humans in Figure 4).

Visually, we observe that most LLMs map uncertainty expressions to probabilities in a consistent way, with higher probabilities for expressions that are perceived by humans as higher-certainty (e.g., "almost certain," "highly likely") and lower probabilities for lower-certainty expressions (e.g., "very unlikely"). Only 2 of the LLMs evaluated, `OLMo (7B)` and `Gemma (2B)`, fail to reproduce this "increasing" pattern across expressions. However, comparatively, the conditional distributions

[6] We focus in this section on a subset of popular models; results for all 10 models evaluated are in Appendix D.



(a) GPT-4o



(b) OLMo-7B Instruct

Figure 5: Distributions of numerical probabilities per uncertainty expression in the non-verifiable setting. Highlighted blue boxes represent the maximum value for each expression.

of LLMs have little variability, tending to be concentrated on only a few probabilities.

These observations are reflected more precisely by the PA scores in Table 1. We observe that larger and newer LLMs (in particular, `GPT-4`, `LLama3 (70B)`, and `Gemini`) perform especially well on this task, matching the modal scores that humans assign to each uncertainty expression. In fact, 8 out of the 10 LLMs evaluated out-perform individual humans, on average, at this task. This aligns with the high-level findings of Maloney et al. (2024), in particular, that the difference between the numerical probabilities of GPT-4 and humans were similar to (or smaller than) inter-human differences. In the context of our experiments, these high scores reflect that LLMs tend to be more consistent than individual humans in terms of agreement with aggregate human responses.

5.2 Does knowledge affect uncertainty perceptions of LLMs?

In this section we assess the extent to which LLMs, and humans, are biased by their prior knowledge or beliefs in mapping uncertainty expressions to numerical probabilities. To investigate this question we collect probability estimates from humans and LLMs on our verifiable (V) dataset, which includes both true and false common-knowledge statements. On average, PA (compared to the non-verifiable responses) for both humans and LLMs is

Table 1: Human-LLM agreement for non-verifiable statements: average Proportional Agreement (PA), PA as a fraction of the *Human Mode* results (% PA), and absolute error between mean responses (MAE). *Human Mode* represents the mode of the human NV distribution, whereas *Human Individual* represents the average behavior across individual humans.

	PA	% PA	MAE
Human Mode	27.6	—	—
Human Individual	17.6	65.9	8.91
ChatGPT	19.7	68.7	6.80
GPT-4	24.4	86.9	4.64
GPT-4o	18.9	68.9	5.58
Gemini	25.4	90.8	4.09
Llama3 (70B)	23.6	85.5	5.56
Mixtral 8x22B	21.8	77.6	7.20

Table 2: Human-LLM agreement for verifiable experiments: average Proportional Agreement (PA), the change in PA from the non-verifiable statements (Table 1) (Δ PA) and absolute error between mean responses (MAE). Again *Human Mode* represents the mode of the human NV distribution, whereas *Human Individual* represents the average behavior across individual humans on the verifiable set.

	PA	Δ PA	MAE
Human Mode	27.6	—	—
Human Individual	16.7	-0.9	9.35
ChatGPT	15.3	-4.4	8.57
GPT-4o	15.2	-3.7	7.05
GPT-4	22.1	-2.3	3.84
Gemini	21.3	-4.1	7.23
Llama3 (70B)	18.9	-4.8	13.73
Mixtral 8x22B	18.6	-3.2	9.78

lower for verifiable statements (Table 2), suggesting that prior knowledge about a statement makes it more difficult to quantify the beliefs of someone else about that statement. This reduction in PA is particularly pronounced for LLMs: all 10 LLMs evaluated demonstrated a reduction in PA, averaging a 4.3 point drop in score, compared to a 0.9 point drop for humans.

To investigate these differences in more detail, we consider the mean numerical probabilities produced by LLMs (Figure 6). These probabilities differ systematically depending on whether the statement is true or false: across the 6 LLMs in Figure 6, the probability generated is on average 7.0 percentage points lower for false than true statements. This indicates that the LLMs' knowledge is "leaking" into the probabilities they produce: the models assign higher numerical probability to the same uncertainty expression when they believe the associated statement it refers to is true than when

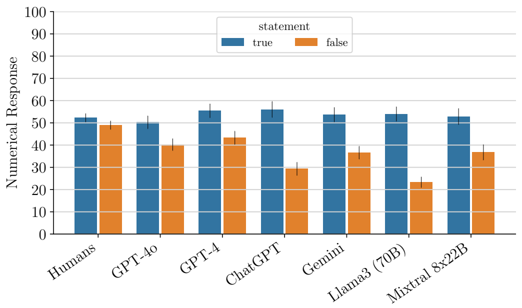


Figure 6: Mean response of the verifiable statements discriminated by truthfulness of statements.

they believe it to be false.

Results for a subset of specific uncertainty expressions are shown in Figure 7. We observe that the prior-knowledge bias differs based on the uncertainty expression: the difference between true and false statements is much higher (49.5 percentage points) for "possible" than for "uncertain", where most models are relatively consistent (averaging a 5.7 percentage point difference).

Overall, we find that all LLMs evaluated demonstrate significant biases based on their prior knowledge, well beyond those of humans. Our results indicate that when an LLM believes a statement is false, it tends to perceive the speaker's certainty as low, regardless of the actual uncertainty expressed by the speaker (and vice versa).

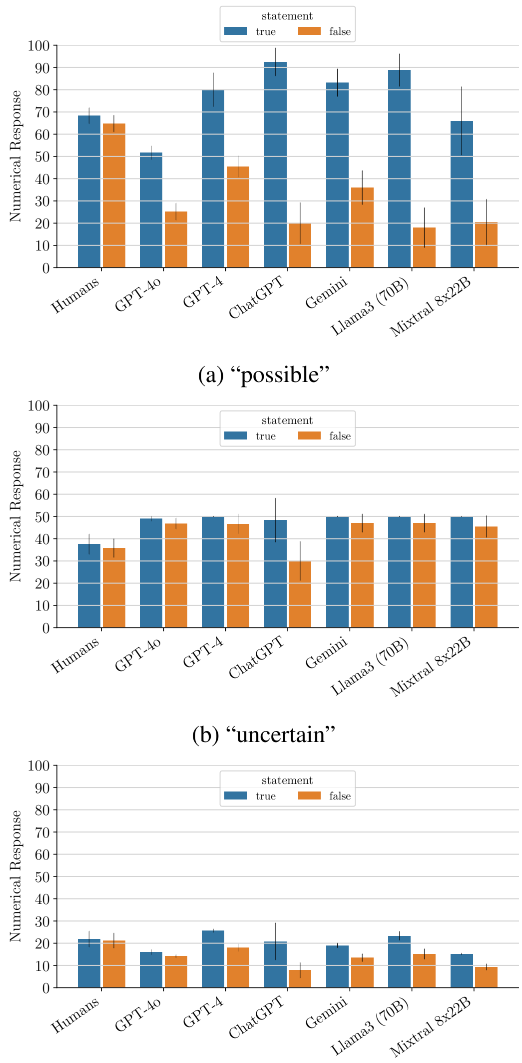
5.3 How generalizable are our findings?

In the previous sections, our analyses are conducted on a manually curated set of 120 statements, comprised of 60 NV statements and 60 V statements. To further validate our findings concerning LLMs' prior knowledge biases, we re-assess the impact of knowledge in LLMs' perceptual capabilities by obtaining their responses for 400 additional verifiable statements. We refer the reader to Appendix E for a more thorough description of the experimental setup. Similarly to the original study, Figure 15 (Appendix) shows that, on average, all models except `Gemma (2B)` exhibit significant perceptual differences between true and false statements—between 5.87 (`OLMo (7B)`) and 17.26 (`LLama3 (70B)`) percentage points. The validation of our verifiable results in this larger dataset corroborates our knowledge bias finding by showing that these perceptual differences persist in a different context.

5.4 How does decoding impact our findings?

The previous analyses employ greedy decoding (i.e., `temperature=0`) when estimating the conditional probability distributions. In this section, we investigate the impact of decoding technique in the model's abilities to perceive linguistic uncertainty by considering richer probability information (i.e., `temperature=1`) during the estimation of the conditional probability distributions[7].

Table 3 summarizes the change in agreement between LLM and human responses between the verifiable and non-verifiable settings (in terms of change in PA and MAE) when using probabilistic decoding. Validating the results reported in Section 5.2 with greedy decoding, we observe a clear



(a) "possible"

(b) "uncertain"

(c) "unlikely"

Figure 7: Mean response for verifiable statements (both true and false) for selected uncertainty expressions.

[7]This analysis requires full probability information, which is prohibitively expensive to obtain empirically through sampling as it would require a large sample size (per (u, s, e)) to faithfully approximate the distribution. As a result, we limit our analysis to OpenAI models for which the top 20 next-token probabilities are available. See Appendix F for additional discussion on this topic.

Table 3: Differences in average proportional agreement and mean responses from non-verifiable to verifiable settings when considering probabilistic decoding (`temperature=1`). Even with a different decoding, we observe the same decrease in LLM perceptions when comparing non-verifiable with verifiable settings.

	Δ PA	Δ MAE
ChatGPT	-3.6	0.4
GPT-4	-3.0	1.9
GPT-4o	-4.2	-0.6

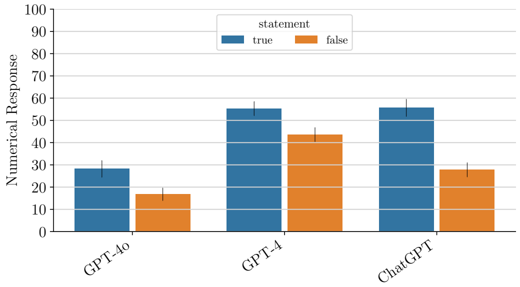


Figure 8: Mean response on the verifiable statements discriminated by truthfulness of statements when decoding probabilistically `temperature=1`.

difference in the PA score between non-verifiable and verifiable statements when using probabilistic decoding. Further, comparing responses across true and false statements, we observe large mean response differences of 11.4, 11.7, and 27.9 percentage points for `GPT-4o`, `GPT-4`, and `ChatGPT`, respectively (see Figure 8 and Figure 17 in the Appendix for a breakdown across expressions). Although `GPT-4o` mean responses are considerably lower than in the greedy decoding setting (i.e., 20 percentage points drop), the gap between true and false statements persists. Ultimately, this analysis confirms the relevance of our previous findings beyond a single decoding strategy.

6 Discussion

Connection to verbalized confidence. Asking LLMs to express their uncertainty through language has become a popular method for obtaining calibrated confidence measures from black-box LLMs (Lin et al., 2022; Tian et al., 2023; Shrivastava et al., 2023; Xiong et al., 2024). Our work raises important questions about the efficacy of this method for text with uncertainty expressions. To date, this technique has not been studied systematically; our findings constitute an initial step in this direction and highlight a need for further research.

Connection to human behavior simulation using LLMs. Our experiments reveal that, despite agreeing with population-level perceptions of linguistic uncertainty, models are not able to capture the full diversity of human behavior. Given the recent interest in using LLMs to simulate human participants (Aher et al., 2023; Gui and Toubia, 2023; Dillion et al., 2023; Park et al., 2023; Namikoshi et al., 2024), our work raises important questions about whose opinions and behaviors are being simulated (Santurkar et al., 2023; Motoki et al., 2023) and reveals a new dimension in which models fail to match the diversity of humans.

Theory of mind. A growing body of work aims to assess the *theory of mind* capabilities of LLMs in different contexts (e.g., Street et al.; Verma et al.; Sap et al.; Zhou et al.). Our task of mapping uncertainty expressions to numerical probabilities, from the perspective of some speaker, is one component of a general theory of mind ability. Our results indicate that LLMs have room for improvement in this area, in particular, that they are prone to confusing their own belief about a statement with the belief of someone else.

7 Conclusions

We introduce the task of assessing the abilities of LLMs to interpret uncertainty in language and evaluate a number of models in this context. We observe that many LLMs can competently map uncertainty expressions to numerical probabilities in a way that aligns with population-level human perceptions, although the probabilities they choose are much less diverse than those by humans. Additionally, we find that LLMs are more susceptible to conflating their own uncertainty about a statement with the statement speaker's uncertainty, resulting in performance that is biased by the LLM's belief about the statement.

In proposing this task, we do not take a stance on whether LLM behavior should mirror the diversity of human behavior—which is a broader philosophical discussion—but focus on characterizing LLMs in comparison to human patterns that arise at the population-level. By highlighting systematic inconsistencies related to the perceptions of linguistic uncertainty in the presence of knowledge, we shed light into overlooked model behaviors that are critical for understanding human-AI communication and downstream LLM performance.

Limitations

US Centric View: In this paper we focus on a small set of uncertainty expressions in English; our baseline is drawn from participants located in the United States. Investigating the role that cultural and language differences play in communicating uncertainty is important future work that will help better characterize the downstream abilities of LLMs for all users.

Lack of Explanation: Our results highlight the LLMs' abilities to interpret uncertainty phrases in a way that agrees with population-level human distribution in the non-verifiable and to less extent in the verifiable setting. We found it surprising to find consistent model performance, especially since we found no evidence of similarly framed tasks in available instruction-tuning and human feedback datasets (Wang et al., 2022; Bai et al., 2022). We hope that future work would explain the reasons behind these findings.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Gati Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Lukas Aichberger, Kajetan Schweighofer, Mykyta Ielanskyi, and Sepp Hochreiter. 2024. Semantically diverse language generation for uncertainty estimation in language models. *Preprint*, arXiv:2406.04306.

Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo,

Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kępa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea

Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirnschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *Preprint*, arXiv:2204.05862.

Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. 2023. Open llm leaderboard. `https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard`.

Wibecke Brun and Karl Halvor Teigen. 1988. Verbal probabilities: Ambiguous, context-dependent, or both? *Organizational Behavior and Human Decision Processes*, 41(3):390–404.

David V. Budescu, Han-Hui Por, Stephen B. Broomell, and Michael Smithson. 2014. The interpretation of ipcc probabilistic statements around the world. *Nature Climate Change*, 4(6):508–512.

Jiuhai Chen and Jonas Mueller. 2023. Quantifying uncertainty in answers from any language model and enhancing their trustworthiness. *Preprint*, arXiv:2308.16175.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv*, abs/1803.05457.

Mandeep K. Dhami and David R. Mandel. 2022. Communicating uncertainty using words and numbers. *Trends in Cognitive Sciences*, 26(6):514–526.

Mandeep K. Dhami and Thomas S. Wallsten. 2005. Interpersonal comparison of subjective probabilities: Toward translating linguistic probabilities. *Memory & Cognition*, 33(6):1057–1068.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can ai language models replace human participants? *Trends in Cognitive Sciences*, 27:597–600.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2024. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, Bangkok, Thailand. Association for Computational Linguistics.

Misty C. Duke. 2023. Probability and confidence: How to improve communication of uncertainty about uncertainty in intelligence analysis. *Journal of Behavioral Decision Making*, 37(1).

Ido Erev and Brent L Cohen. 1990. Verbal versus numerical probabilities: Efficiency, biases, and the preference paradox. *Organizational Behavior and Human Decision Processes*, 45(1):1–18.

Wade Fagen-Ulmschneider. 2019. Perception of probability words. Accessed: [June 12, 2024].

Joe Fore. 2019. "a court would likely (60-75%) find...." defining verbal probability expressions in predictive legal analysis. *Legal Comm. & Rhetoric: JAWLD*, 16:49.

George Gui and Olivier Toubia. 2023. The challenge of using llms to simulate human behavior: A causal inference perspective. *ArXiv*, abs/2312.15524.

Emily H. Ho, David V. Budescu, Mandeep K. Dhami, and David R. Mandel. 2015. Improving the communication of uncertainty in climate science and intelligence analysis. *Behavioral Science & Policy*, 1(2):43–55.

Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang, and Yang Zhang. 2024. Decomposing uncertainty for large language models through input clarification ensembling.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Tzur M. Karelitz and David V. Budescu. 2004. You say “probable” and i say “likely”: Improving interpersonal communication with verbal probability phrases. *Journal of Experimental Psychology: Applied*, 10(1):25–41.

Sunnie S. Y. Kim, Q. Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. "i’m not sure, but...": Examining the impact of large language models’ uncertainty expression on user reliance and trust. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’24, page 822–835, New York, NY, USA. Association for Computing Machinery.

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, pages 12–24.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *Proceedings of the 11th International Conference on Learning Representations*, ICLR’23.

Weixin Liang, Girmaw Abebe Tadesse, Daniel Ho, L. Fei-Fei, Matei Zaharia, Ce Zhang, and James Zou. 2022. Advances, challenges and opportunities in creating data for trustworthy AI. *Nature Machine Intelligence*, 4(8):669–677.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *Preprint*, arXiv:2205.14334.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024. Generating with confidence: Uncertainty quantification for black-box large language models. *Preprint*, arXiv:2305.19187.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. *arXiv preprint arXiv:2109.05052*.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, et al. 2023. A pretrainer’s guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. *arXiv preprint arXiv:2305.13169*.

Laurence T. Maloney, Maria F. Dal Martello, Vivian Fei, and Valerie Ma. 2024. A comparison of human and gpt-4 use of probabilistic phrases in a coordination game. *Scientific Reports*, 14(1).

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.

Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. 2023. More human than human: measuring chatgpt political bias. *Public Choice*, 198(1–2):3–23.

Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. 2024. More human than human: Measuring chatgpt political bias. *Public Choice*, 198(1):3–23.

Keiichi Namikoshi, Alexandre L. S. Filipowicz, David A. Shamma, Rumen Iliev, Candice Hogan, and Nikos Aréchiga. 2024. Using llms to model the beliefs and preferences of targeted populations. *ArXiv*, abs/2403.20252.

Joon Sung Park, Joseph O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST ’23, New York, NY, USA. Association for Computing Machinery.

Anthony Patt and Suraje Dessai. 2005. Communicating uncertainty: lessons learned and suggestions for climate change assessment. *Comptes rendus. Géoscience*, 337(4):425–441.

Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. 2024. In-context impersonation reveals large language models’ strengths and biases. *Advances in Neural Information Processing Systems*, 36.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? *Preprint*, arXiv:2303.17548.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large lms. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3762–3780.

Ruta Sawant and Sujit Sansgiry. 2018. Communicating risk of medication side-effects: role of communication format on risk perception. *Pharmacy Practice*, 16(2):1174.

Nino Scherrer, Claudia Shi, Amir Feder, and David Blei. 2024. Evaluating the moral beliefs encoded in llms. *Advances in Neural Information Processing Systems*, 36.

Nikil Selvam, Sunipa Dev, Daniel Khashabi, Tushar Khot, and Kai-Wei Chang. 2023. The tail wagging the dog: Dataset construction biases of social bias benchmarks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1373–1386, Toronto, Canada. Association for Computational Linguistics.

Preethi Seshadri, Pouya Pezeshkpour, and Sameer Singh. 2022. Quantifying social biases using templates is unreliable. *Preprint*, arXiv:2210.04337.

Vaishnavi Shrivastava, Percy Liang, and Ananya Kumar. 2023. Llamas know what gpts don't show: Surrogate models for confidence estimation. *ArXiv*, abs/2311.08877.

Chenglei Si, Chen Zhao, Sewon Min, and Jordan Boyd-Graber. 2022. Re-examining calibration: The case of question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2814–2829, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Aaditya K. Singh and DJ Strouse. 2024. Tokenization counts: the impact of tokenization on arithmetic in frontier llms. *Preprint*, arXiv:2402.14903.

Mark Steyvers, Heliodoro Tejeda, Aakriti Kumar, Catarina Belem, Sheer Karny, Xinyue Hu, Lukas Mayer, and Padhraic Smyth. 2024. The calibration gap between model and human confidence in large language models. *Preprint*, arXiv:2401.13835.

Winnie Street, John Oliver Siy, Geoff Keeling, Adrien Baranes, Benjamin Barnett, Michael McKibben, Tatenda Kanyere, Alison Lentz, Robin IM Dunbar, et al. 2024. Llms achieve adult human performance on higher-order theory of mind tasks. *arXiv preprint arXiv:2405.18870*.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher 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, Singapore. Association for Computational Linguistics.

Mudit Verma, Siddhant Bhambri, and Subbarao Kambhampati. 2024. Theory of mind abilities of large language models in human-robot interaction: An illusion? In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '24, page 36–45, New York, NY, USA. Association for Computing Machinery.

Thomas S Wallsten, David V Budescu, Amnon Rapoport, Rami Zwick, and Barbara Forsyth. 1986a. Measuring the vague meanings of probability terms. *Journal of Experimental Psychology: General*, 115(4):348.

Thomas S. Wallsten, David V. Budescu, Rami Zwick, and Steven M. Kemp. 1993. Preferences and reasons for communicating probabilistic information in verbal or numerical terms. *Bulletin of the Psychonomic Society*, 31(2):135–138.

Thomas S Wallsten, Samuel Fillenbaum, and James A Cox. 1986b. Base rate effects on the interpretations of probability and frequency expressions. *Journal of Memory and Language*, 25(5):571–587.

Thomas S. Wallsten, Yaron Shlomi, and Hisuchi Ting. 2008. Exploring intelligence analysts' selection and interpretation of probability terms: Final report for research contract 'expressing probability in intelligence analysis'.

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. "kelly is a warm person, Joseph is a role model": Gender biases in llm-generated reference letters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3730–3748.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Maitreya Patel, Kuntal Kumar Pal, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Shailaja Keyur Sampat, Savan Doshi, Siddhartha Mishra, Sujan Reddy, Sumanta Patro, Tanay Dixit, Xudong Shen, Chitta Baral, Yejin Choi, Noah A. Smith, Hannaneh Hajishirzi, and Daniel Khashabi. 2022. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. *Preprint*, arXiv:2204.07705.

S. J. W. Willems, Casper Johannes Albers, and Ionica Smeets. 2019. Variability in the interpretation of dutch probability phrases - a risk for miscommunication. *arXiv: Other Statistics*.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *Proceedings of the 12th International Conference on Learning Representations*, ICLR'24.

Kaitlyn Zhou, Jena D. Hwang, Xiang Ren, and Maarten Sap. 2024. Relying on the unreliable: The impact of language models' reluctance to express uncertainty. *ArXiv*, abs/2401.06730.

Pei Zhou, Aman Madaan, Srividya Pranavi Potharaju, Aditya Gupta, Kevin R McKee, Ari Holtzman, Jay Pujara, Xiang Ren, Swaroop Mishra, Aida Nematzadeh, et al. 2023. How far are large language models from agents with theory-of-mind? *arXiv preprint arXiv:2310.03051*.

Jian-Qiao Zhu and Thomas L. Griffiths. 2024. Incoherent probability judgments in large language models. *Preprint*, arXiv:2401.16646.

A Human Experiments

Human responses were collected using Prolific (`https://www.prolific.com/`). We recruited 100 participants for the non-verifiable experiment and 100 different participants for the second verifiable experiment. One of the 100 responses was not received due to a technical issue in both the first and second experiment, leaving a total of 99 responses for each. We recruited participants whose first language was English that were located in the United States. Participants were paid $2 for completing the study and the average completion time was 8 minutes and 48 sections; the average payment rate was $13.64/hour. The University of California, Irvine Institutional Review Board (IRB) approved the experimental protocol. Prior to the experiment, participants were given detailed instructions outlining the experimental procedure as well as how to understand and interact with the user interface. Participants were asked to sign an integrity pledge after reading all of the instructions, stating that they would complete the experiment to the best of their abilities. After submitting their integrity pledge, participants were granted access to the experiment.

We filtered out low-quality responses with the following procedure. For each participant, we computed the Spearman correlation between the participant's responses and the overall ranking of uncertainty statements in the non-verifiable experiment. We removed participants with $\rho < 0.2$, a threshold chosen empirically to filter out only no-signal, spam-like responses. This filter removed 5 participants in the first experiment and 10 in the second experiment. The final totals are 94 participants in the non-verifiable experiment and 89 in the verifiable experiment.

B Greedy Samples

Gemini pro directly produces numbers (14 unique answers).

For Llama-70B sample greedy, we determine that 95% of the responses produce numerical values between 0 and 100. For the remaining 47 examples, the first produced number in the response is the proposed model answer in about 46.

Despite more verbose, DBRX-Instruct responses also start with the number and then proceed to explain its reasoning. In 5.55% of examples (50 out of 900), DBRX-Instruct proposes an enumerate of numbers (e.g., "0, 5, or 10." expressed for uncertainty expressions like 'very unlikely' and 'highly unlikely').

C Experiment Details

In this section, we provide the details concerning the different aspects of the experiments carried in this paper.

C.1 Uncertainty Expressions

The uncertainty expressions are a subset of the expressions proposed in Wallsten et al.; Wallsten et al.; Willems et al.; Fore. The final list of uncertainty expressions used in this paper is listed below:

1. almost certain
2. highly likely
3. very likely
4. likely
5. probable
6. somewhat likely
7. somewhat unlikely
8. uncertain
9. possible
10. unlikely
11. not likely
12. doubtful
13. very unlikely
14. highly unlikely

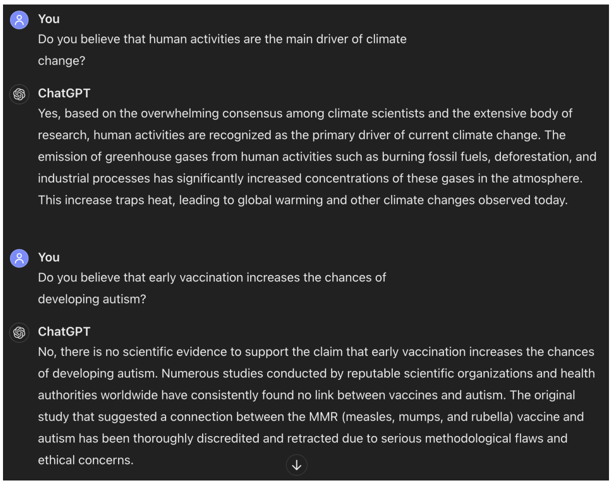


Figure 9: Responses from ChatGPT when asked about its belief in the statements from Figure 1. ChatGPT agrees that "human activities are the main driver of climate change," but disagrees with the statement that "early vaccination increases the chances of developing autism."

Gender	List of names
Female	"Amanda", "Bonnie", "Camille", "Catherine", "Cheri", "Ethel", "Gabriela", "Jacquelyn", "Jessica", "Laura", "Olga", "Roxanne", "Silvia", "Tara", "Violet"
Male	"Brendan", "Bruce", "David", "Gary", "Isaac", "Jeffery", "Joey", "Johnnie", "Kenny", "Lance", "Marco", "Mike", "Nathan", "Nick", "Raul"

Table 4: Names used in the experiments.

C.2 Name Selection

Table 4 lists all names used in our experiment, discriminated by gender. The names are generated using a random name generator[8], configured to generate 32 names from the United States (see Table 4). We iteratively generate names until we obtain 16 unique male names and 16 unique female names.

C.3 Prompts

In our main paper, we conduct experiments using 2 demonstrations. This guarantees tight similarity between the setup employed in the study of liguinstic perceptions of LLMs and humans.

During the course of our experiments, we carried experiments with varying assumptions: non-verifiable setup assessed models (and humans) perceptions in the absence of strong prior beliefs about the events, whereas the verifiable setup focused on the evaluation of the same perceptions when knowledge was present. We used two different sets of exemplars in our experiments to reflect the different setups.

For the **non-verifiable experiments**, we used the following uncertainty expression, speaker, statement pairs:

- `speaker`: "Kathleen", `uncertainty`: "impossible", `statement`: "the cafe made a profit in the last 6 months".
- `speaker`: "Cedric", `uncertainty`: "certain", `statement`: "the new treatment will improve the patient's condition".

For the **verifiable experiments**, we used the following uncertainty expression, speaker, statement pairs:

- `speaker`: "Kathleen", `uncertainty`: "impossible", `statement`: "the Sun orbits around the planet Earth".
- `speaker`: "Cedric", `uncertainty`: "certain", `statement`: "all metals can conduct electricity".

C.4 Non-Verifiable Statements

In this section, we list a random selection of 20 statements (5 per each statement type). These statements are incorporated in the prompts listed in the

[8] https://randomwordgenerator.com/name.php, last accessed on March 26th, 2024.

In this experiment, you will be shown text reflecting an individual's beliefs about a statement. You will then be asked to judge, in your opinion, the strength of that individual's belief. To do so, you will provide a probability (expressed as a percentage) where:
- The number 0 indicates that the individual believes with certainty that the statement is false.
- The number 100 indicates that the individual believes with certainty that the statement is true.

You will be shown two example question and response pairs below to familiarize you with the experiment setup. After the examples, you will be asked to answer a new question.

Example Question 1
`[[demonstration1]]`
Choose the answer from the following options:
0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100
Correct answer: 100

Example Question 2
[[demonstration2]]|
Choose the answer from the following options:
0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100
Correct answer: 100

Question
Given the examples before, answer the following question by writing a single number as the answer.

`[[test-example]]`
Choose the answer from the following options:
0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100 Correct answer:

Figure 10: Prompt used to conduct the experiments in the 2-shot setting. `[[demonstration1]]`, `[[demonstration2]]`, and `[[text-example]]` are placeholders that will be composed of a statement (or event), question, and individual instructions. The prompt closely reflects the interface presented to the participants of the user studies. Similarly to the human interface, we encourage LLMs to select a multiple of 5 in the range [0, 100].

Sentence: "[[speaker]] believes it's [[uncertainty]] that [[statement]]."

Question: Provide a number between 0 and 100 that reflects, from [[speaker]]'s perspective, the probability (expressed as a percentage) that [[statement]].
- 0 indicates [[speaker]] believes there is a 0% chance that [[statement]].
- 100 indicates [[speaker]] believes there is a 100% chance that [[statement]].

Figure 11: Structure of a demonstration, where the placeholders [[speaker]] are replaced by gendered names, [[uncertainty]] is replaced by the uncertainty expressions, and [[statement]] is replaced with the corresponding statements.

main paper and the placeholders [[they]] and [[their]] are replaced by pronouns matching the gender of the statement speaker's name.

Forecasting of future events. Verbal probabilities are often used to communicate uncertainty about future events.

1. [[they]] will buy a new watch this Thanksgiving weekend.
2. [[they]] will be offered a promotion this fall.
3. the company will have another round of layoffs by mid July.
4. there will be vegetarian options at the barbecue.
5. [[they]] will visit New York over winter break.

Imperfect knowledge. Verbal probabilities can also be used to communicate uncertainty imprecise information about events or outcomes.

1. the restaurant near [[their]] apartment accepts reservations.
2. the new museum is offering complimentary admission.
3. there is a yoga studio within 2 miles of [[their]] workplace.
4. there are more than eighty students in the auditorium right now.
5. the temperature in the office is at least 72 degrees Fahrenheit.

Possession. Alternatively, verbal probabilities can be used to convey uncertainty about acquaintances, be it in terms of the objects they own or in terms of their preferences.

1. [[their]] boss owns a blue car.
2. [[their]] friend has a leather jacket.
3. [[their]] cousin has a vegetable garden.
4. [[their]] classmate owns a guitar.
5. [[their]] boss has a stereo amplifier.

Preference.

1. [[their]] cousin prefers spinach over broccoli.
2. [[their]] boss prefers coffee over tea.
3. [[their]] friend prefers running over cycling.
4. [[their]] neighbor prefers the beach over the mountains.
5. [[their]] coworker prefers reading books over watching movies.

C.5 Verifiable Statements

In this section, we list a random selection of 9 true statements (3 per each topic) and their false counterparts. These statements are incorporated then used as part of the test examples the prompts listed in Section C.3.

Geography. One of the topics of the experiment involves geography, as well as knowledge about landmarks and monuments. These statements were curated from a set of easy trivia questions provided by The Question Company (as described in Section 4). For each question-answer pair in the trivia dataset, we create both a true and a false statement using the correct and one incorrect answer choice, respectively.

Given our interest in attesting the knowledge effect in the models' and humans' perceptions of linguistic uncertainty, we purposely decided to use easy trivia questions as the basis for our facts (as opposed to more difficult ones), since this subset constitutes a good proxy for facts that LLMs and humans may have strong prior beliefs about.

1. Great Britain directly borders 0 countries.
2. the Colosseum, a famous landmark in Rome, was originally built as an Amphitheatre.
3. New York is known as the Big Apple.
4. Great Britain directly borders 2 countries.
5. the Colosseum, a famous landmark in Rome, was originally built as an Cathedral.
6. New York is known as the Big Orange.

History of Art. One of the topics of the experiment involves history of arts. For each fact we include both a true and one false variation of that fact.

1. the Mona Lisa is a famous painting by Leonardo da Vinci.
2. the Scream is the best known painting by Edvard Munch.
3. Andy Warhol became a famous artist in the 1960s for painting soup cans and soap boxes.
4. the Mona Lisa is a famous painting by Tintoretto.
5. the Scream is the best known painting by Jackson Pollock.
6. Frida Kahlo became a famous artist in the 1960s for painting soup cans and soap boxes.

Science. These include facts concerning chemistry, biology, and astronomy. For each fact we include both a true and one false variation of that fact.

1. water's chemical formula is H2O.
2. pH is a measure of the acidity or basicity of a substance.
3. the nearest planet to the sun is Mercury.
4. carbon monoxide's chemical formula is H2O.
5. oG is a measure of the acidity or basicity of a substance.
6. the nearest planet to the sun is Mars.

C.6 Language Models

Throughout our paper, we use OpenAI to obtain the results for `ChatGPT`, `GPT-4`, and `GPT-4o`; Google's Vertex AI APIs to obtain results for `Gemini`, TogetherAI[9] to run `LLama3 (70B)`, `Mixtral 8x7B`, `Mixtral 8x22B`, and DBRX. We run `LLama3 (8B)` `OLMo (7B)` and `Gemma (2B)` locally on a single GPU 8 RTX A6000 (48 GB).

During the paper, we shorten the name of the studied models for simplicity. All our experiments consider the instruction-tuned or RLHF version of the mentioned models. All experiments were conducted from April through June. For reproducibility, we list below the mapping from model name to exact version of the model used:

- `ChatGPT`: `gpt-3.5-turbo-0125`
- `GPT-4`: `gpt-4-turbo-2024-04-09`
- `GPT-4o`: `gpt-4o-2024-05-13`
- `LLama3 (8B)`: `meta-llama/Meta-Llama-3-8B-Instruct`
- `LLama3 (70B)`: `meta-llama/Llama-3-70b-chat-hf`
- `Gemini`: `models/gemini-pro`
- `Mixtral 8x7B`: `mistralai/Mixtral-8x7B-Instruct-v0.1`
- `Mixtral 8x22B`: `mistralai/Mixtral-8x22B-Instruct-v0.1`
- `Gemma (2B)`: `google/gemma-1.1-2b-it` (we found `Gemma (2B)` to respond better empirically to the prompts than its 7B version, which tended to extrapolate the few-shot instructions with additional examples).
- `OLMo (7B)`: `allenai/OLMo-7B-Instruct`

In Section 4.2 we describe the use of different methodologies to extract numerical responses from LLMs. The following describes the list of methodologies used for each LLM:

[9] https://www.together.ai/

- Greedy Sampling: `Gemini`, `LLama3 (70B)`, `Mixtral 8x7B`, `Mixtral 8x22B`. We opt for using greedy sampling as opposed to standard sampling due to budget constraints. Given the nature of our experiments, faithfully estimating the empirical distributions over the 100 numbers would require hundreds or thousands of calls. These calls are time-consuming and costly. We believe that using decoding is still representative of how a model would behave in most cases.

- Full next-token probability distribution: `LLama3 (8B)`, `Gemma (2B)`, `OLMo (7B)`. We found these models to be particularly brittle to the prompts.

- Next-token probability distribution: `ChatGPT`, `GPT-4`, `GPT-4o`. As of June 2024, OpenAI models only provide access to the next-token probabilities of the top-20 tokens. During the experiments in Section 5.4, we collect the information about the top 20 numbers

C.6.1 Full Next-Token Probability Information

By definition, our task elicits a numerical response from LLMs, which resembles the setup in verbalized confidence (Tian et al., 2023). The adoption of single digit tokenization (Singh and Strouse, 2024) by autoregressive models (e.g., `Gemma (2B)`, `LLama3 (70B)`, and `OLMo (7B)`) creates some challenges in the computation of numerical responses non-trivial for autoregressive models. In practice, due to the left-to-right nature of LLMs, single digit tokenization implies that the probability of a number between $[0, 9]$ is always greater or equal to the probability of any number in $[10, 100]$. To circumvent this problem, we report the *corrected probability* during our experiments as follows:

$$p_{\text{model}}(y_t = i|x) - \sum_{j=0}^{9} p_{\text{model}}(y_t = i, y_{t+1} = j|x)$$

, where x is a prompt and $j \in [0, 9]$. Intuitively, this means that we are computing the probability of $i \in [0, 9]$ and no other number following it. The details of what a number is change with tokenizer implementation.

Selecting the greedy prediction: Unlike traditional greedy decoding, we condition the selection of the arg-max prediction to the set of strings representing the numbers between $[0, 100]$ (followed by no other number).

C.6.2 Partial Next-Token Probability Information

In order to work, this method requires two properties to be satisfied: (1) numbers between 0 and 100 were encoded with unique tokens (i.e., there are 101 unique integers that represent each individual token), and (2) exponentiating the log probabilities returned by the black-box API must lead to a valid probability distribution (i.e., numbers obtained for different prompts will be comparable to one another). For OpenAI models, the first requirement is satisfied.

Selecting the greedy prediction: Unlike traditional greedy decoding, we condition the selection of the arg-max prediction over numbers the top-k (k=20 for OpenAI). That is, we select the most likely number that is present in the top-20 predicted tokens.

Estimating conditional distributions using probabilistic decoding: In Section 5.4, we use the information available in the top-20 tokens made available by OpenAI models. Like before we bin the predictions into 21 bins, defined from 0 to 100 in increments of 5. To ensure that we have a valid probability definition, we consider an additional bin ("-1") that accumulates the probability of not having being a number in the top-20.

D Additional Results

In this section, we report additional results, including visualization of the empirical distributions for models and humans in Section D.1, additional measurements of similarity between non-verifiable and verifiable distributions in Section D.2, metrics discriminated by uncertainty expression (whenever applicable) in Sections D.3 and D.4.

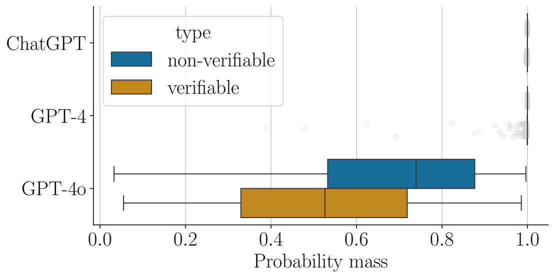
D.1 Histograms

Figure 13 depicts the empirical distributions for the non-verifiable experiments.

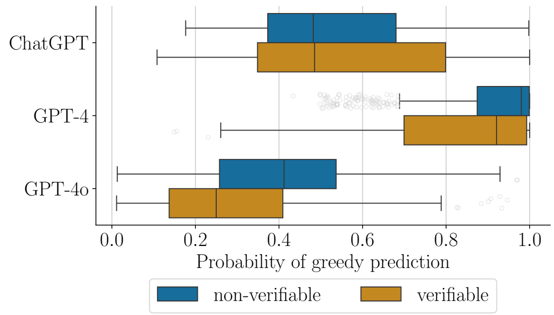
D.2 Summary Metrics

For a more complete understanding of the differences among the distributions, we report distance metrics in Table 5.

- **Proportional Agreement**: proposed in Section 4.3, measures the overall agreement between an agent's and a reference (population) distribution. We use the results of the human



(a) Total probability mass assigned to a number across models and settings.



(b) Probability mass of greedy prediction.

Figure 12: Differences in the probability mass as determined by OpenAI models on the top-20 tokens. We report these values across all statements (n=840) in the verifiable and non-verifiable settings. We observe that numbers account for the majority of probability mass in `ChatGPT` and `GPT-4`. Upon analysis we found lower probability mass assigned by `GPT-4` to be correlated with the appearance of the words "Given", "The", and "And".

studies in the non-verifiable setting as our reference distribution throughout the whole paper.

- **Mean Absolute Error**: proposed in Section 4.3, measures the average agreement across uncertainty expressions between a agent's distribution and a reference (population) distribution. In this case, we also use the human results from the non-verifiable setting as our reference distribution throughout the paper.

- **Wasserstein Distance**: computed using `scipy.stats.wasserstein_distance`, measures the distance between two conditional distributions.

D.3 Proportional Agreement

Tables 6 and 7 report the proportional agreement (PA) metric discriminated by uncertainty expression in the non-verifiable and verifiable settings, respectively. The results are reported in the filtered pool of human participants.

D.4 Mean Response

Figure 14 illustrates the mean rated probability metric discriminated by uncertainty expressions across the non-verifiable, as well as the true and false verifiable statements.

E Generalization results

Diversity of grammatical and semantic structures is an important component of current evaluation practices in LLMs (Selvam et al., 2023; Seshadri et al., 2022), since it helps ensure that obtained results are not an artifact of the evaluation methodology and/or benchmarks used. The experiments described in the main paper were carefully crafted to cover various topics and situations where uncertainty expressions could be used. To further strengthen our analysis and validate our findings, we simultaneously run collect models perceptions of uncertainty expressions using a larger dataset. This dataset by the authors based on the AI2-Arc test set (Clark et al., 2018) — a popular question-answering dataset consisting of genuine grade-school level, multiple-choice science questions. Not only has this dataset been recently used to measure commonsense reasoning of current state-of-the-art LLMs (Jiang et al., 2023; Achiam et al., 2024; Beeching et al., 2023), but it is also composed of easier questions, a key aspect to our verifiable experiment setup.

The creation of this dataset mirrors the procedure described in Section 4. We manually repurposed 200 question-answer pairs from AI2 Arc (100 from the easy set and another 100 from the challenge set). For every statement, the authors produce a true statement and a false statement using the available information about the correct and incorrect multiple choices. The final dataset consists of 200 true statements and 200 false statements.

To determine distributional differences between the conditional distributions obtained in the main paper and the ones obtained in the generalization set, we compare the Wasserstein-1 distance of the two empirical distributions. These values are reported in Table 8. In general, we find models that performed worse in the main paper, including `Gemma (2B)` and `OLMo (7B)`, to exhibit the largest distributional differences with Wasserstein distances of 48.5 and 13.1 when averaged over uncertainty expressions. `ChatGPT`, `LLama3 (70B)`,

Table 5: Summary metrics averaged across uncertainty expressions. All metrics are computed with respect to the human distribution in the non-verifiable setting. “PA” reports the general agreement between LLMs and the mode of the human distribution, reported in percentages. “MAE” reports the absolute error between the mean responses of LLMs and those of humans. Wasserstein-1 computes the distance between LLMs and human distributions.

		Avg PA (↑)		Avg MAE (↓)		Avg Wasserstein-1 (↓)	
		NV	V	NV	V	NV	V
Human	Mode	27.6	27.6	—	—	—	—
	Individual	17.6	16.7	8.91	9.35	12.35	12.99
Baseline	Random	5.1	5.1	27.72	27.72	28.16	28.16
LLM	OLMo	12.1	7.6	18.44	33.67	20.45	40.16
	Gemma (2B)	8.1	6.6	20.17	24.33	22.13	25.89
	Llama3 8B	17.8	10.1	11.99	16.59	14.11	18.35
	Llama3 (70B)	23.6	18.8	5.56	13.73	9.94	16.39
	Mixtral 8x7B	21.8	15.2	5.88	12.32	8.88	15.93
	Mixtral 8x22B	21.8	18.6	7.20	9.78	10.78	12.05
	Gemini	25.4	21.3	4.09	7.23	9.24	9.78
	ChatGPT	19.7	15.3	6.80	8.57	9.26	12.74
	GPT-4	24.4	22.1	4.64	3.84	9.96	6.88
	GPT-4o	18.9	15.2	5.58	7.05	10.34	9.96

and `Mixtral` models all exhibit higher differences in expressions of higher certainty, e.g., “highly likely”, “probable”, “possible”. On the other hand, the two `GPT-4` models, as well as Gemini (Pro) suffer the least changes distributionally (1.9, 1.8, and 4.2 Wasserstein-1 distances on average, respectively), suggesting that these models were robust to changes in the statements.

In the main paper, we find it surprising that LLMs perception abilities differ significantly based on whether the uncertainty expressions are referring to someone’s belief in a true or false statement. To test the generalization of this finding in a larger (and different) dataset, we repeat the same analysis and compare the observed mean response differences with that of humans obtained in the original setting (see Figures 15 and 16). We observe that in absolute sense the differences are smaller than those observed in the original setting, but that models are affected by this knowledge gap to a greater extent than humans.

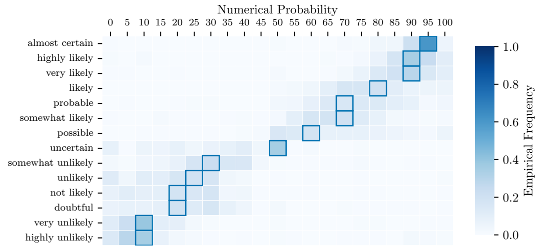
F Probabilistic decoding

Table 6: Proportional Agreement (PA) of models in the non-verifiable setting.

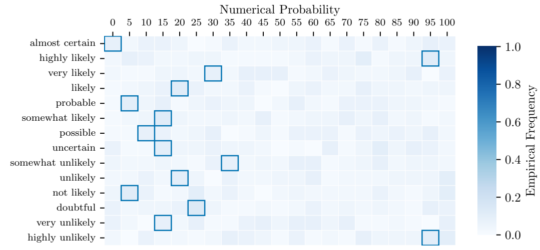
	Humans Mode	Humans Ind	GPT-4o	GPT-4	ChatGPT	Llama3	Mixtral 8x22B	Gemini	DBRX	Rand
Avg PA	27.6	17.6	18.9	24.4	19.7	23.6	21.8	25.4	14.2	5.1
almost certain	60.6	42.0	35.5	60.6	55.9	58.7	60.6	60.6	25.9	7.0
highly likely	34.6	22.5	17.4	22.1	16.1	8.3	23.9	30.0	7.8	6.5
very likely	28.7	17.7	19.1	19.5	13.6	14.4	15.1	21.9	14.8	5.1
probable	16.0	11.1	14.9	14.3	7.6	14.8	15.7	14.1	10.3	5.1
likely	20.2	12.9	16.5	16.5	8.2	19.7	16.5	18.1	8.7	5.0
somewhat likely	20.2	13.5	12.8	17.9	6.8	16.5	18.2	16.2	6.3	5.7
somewhat unlikely	22.3	14.0	18.0	19.3	21.8	22.3	18.4	19.9	15.6	4.9
uncertain	35.1	16.9	35.1	35.1	33.1	35.1	35.1	35.1	35.1	3.4
possible	18.1	10.7	6.5	15.4	14.2	15.0	8.6	15.5	15.6	5.1
unlikely	19.1	13.5	11.1	18.3	15.9	16.8	10.1	16.5	14.1	4.8
not likely	18.1	12.6	11.3	17.1	16.7	17.7	16.1	18.1	10.7	5.7
doubtful	19.7	11.7	14.2	12.9	13.2	17.5	16.4	19.7	11.8	5.8
very unlikely	38.8	23.4	27.8	37.9	26.5	38.8	22.3	36.9	11.5	3.0
highly unlikely	35.1	23.5	24.8	35.1	25.7	35.1	28.7	32.8	10.9	3.8

Table 7: Proportional Agreement (PA) of models in the verifiable setting.

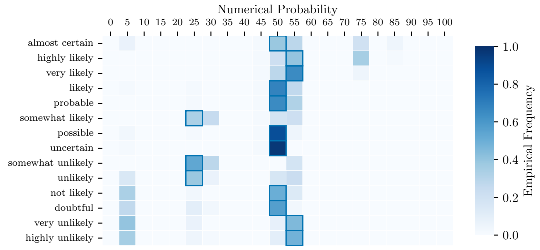
	Humans Mode	Humans Ind	GPT-4o	GPT-4	GPT-3.5	Llama3	Mixtral 8x22B	Gemini	DBRX	Rand
Avg PA	27.6	16.7	15.2	22.1	15.3	18.8	18.6	21.3	13.6	5.1
almost certain	60.6	39.7	28.7	60.6	48.0	35.2	60.6	54.6	24.2	7.0
highly likely	34.6	21.5	14.7	22.0	13.0	17.4	24.1	22.8	15.5	6.5
very likely	28.7	17.7	17.2	17.2	10.1	14.9	14.9	17.6	12.7	5.1
probable	16.0	11.4	11.3	11.5	3.8	6.8	5.8	9.9	6.6	5.1
likely	20.2	12.7	12.7	12.4	7.3	10.7	9.2	11.3	6.4	5.0
somewhat likely	20.2	13.3	7.0	12.1	6.5	8.6	9.0	10.0	7.6	5.7
somewhat unlikely	22.3	13.0	15.0	18.8	15.4	17.3	11.8	15.4	14.7	4.9
uncertain	35.1	15.9	30.3	34.2	25.1	34.0	33.4	34.0	31.8	3.4
possible	18.1	9.4	6.1	10.2	5.8	4.0	4.6	6.7	6.9	5.1
unlikely	19.1	13.2	10.6	16.8	13.6	15.0	9.4	14.3	12.6	4.8
not likely	18.1	11.7	10.9	15.0	12.8	15.5	13.3	15.7	12.7	5.7
doubtful	19.7	10.0	11.6	11.5	6.2	15.2	13.8	16.4	11.9	5.8
very unlikely	38.8	21.8	18.8	34.1	24.2	35.9	22.3	36.2	12.8	3.0
highly unlikely	35.1	22.7	17.2	32.8	22.1	33.3	28.7	33.3	14.0	3.8



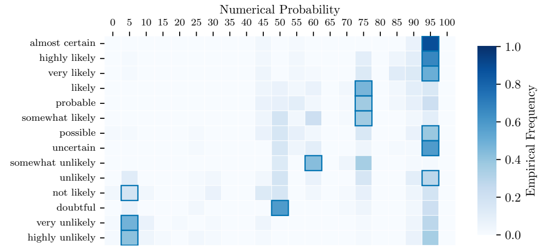
(a) Humans



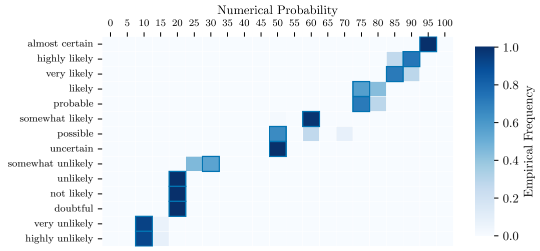
(b) Random Baseline



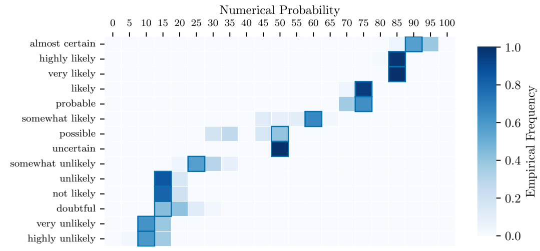
(c) `Gemma-1.1-2b-it`



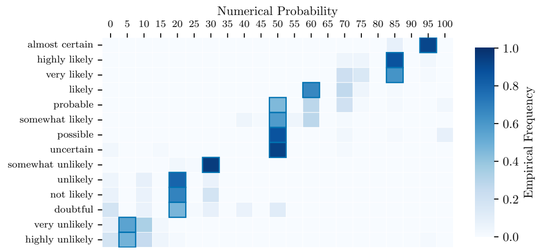
(d) `OLMo-7B-instruct`



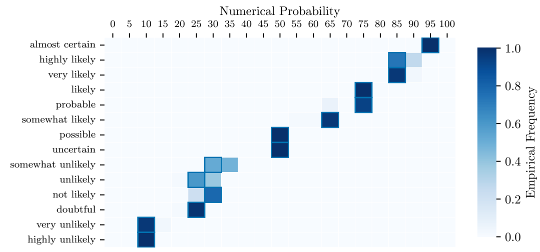
(e) `gemini-pro`



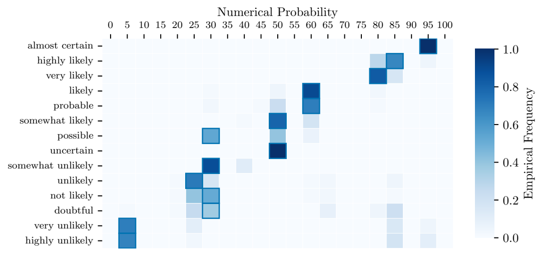
(f) `GPT-4o`



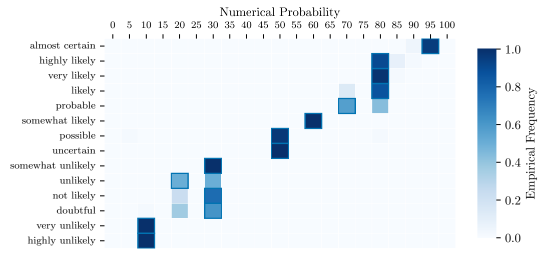
(g) `ChatGPT`



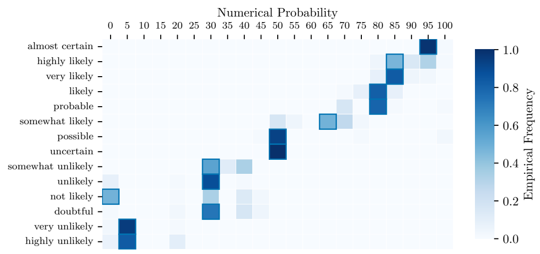
(h) `GPT-4`



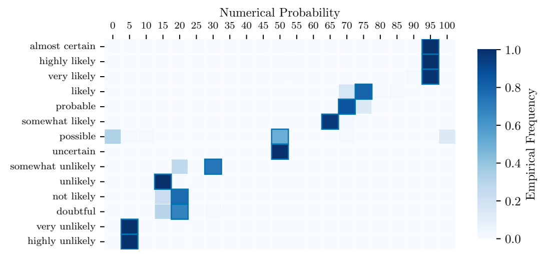
(i) `LLama3 (8B)`



(j) `LLama3 (70B)`

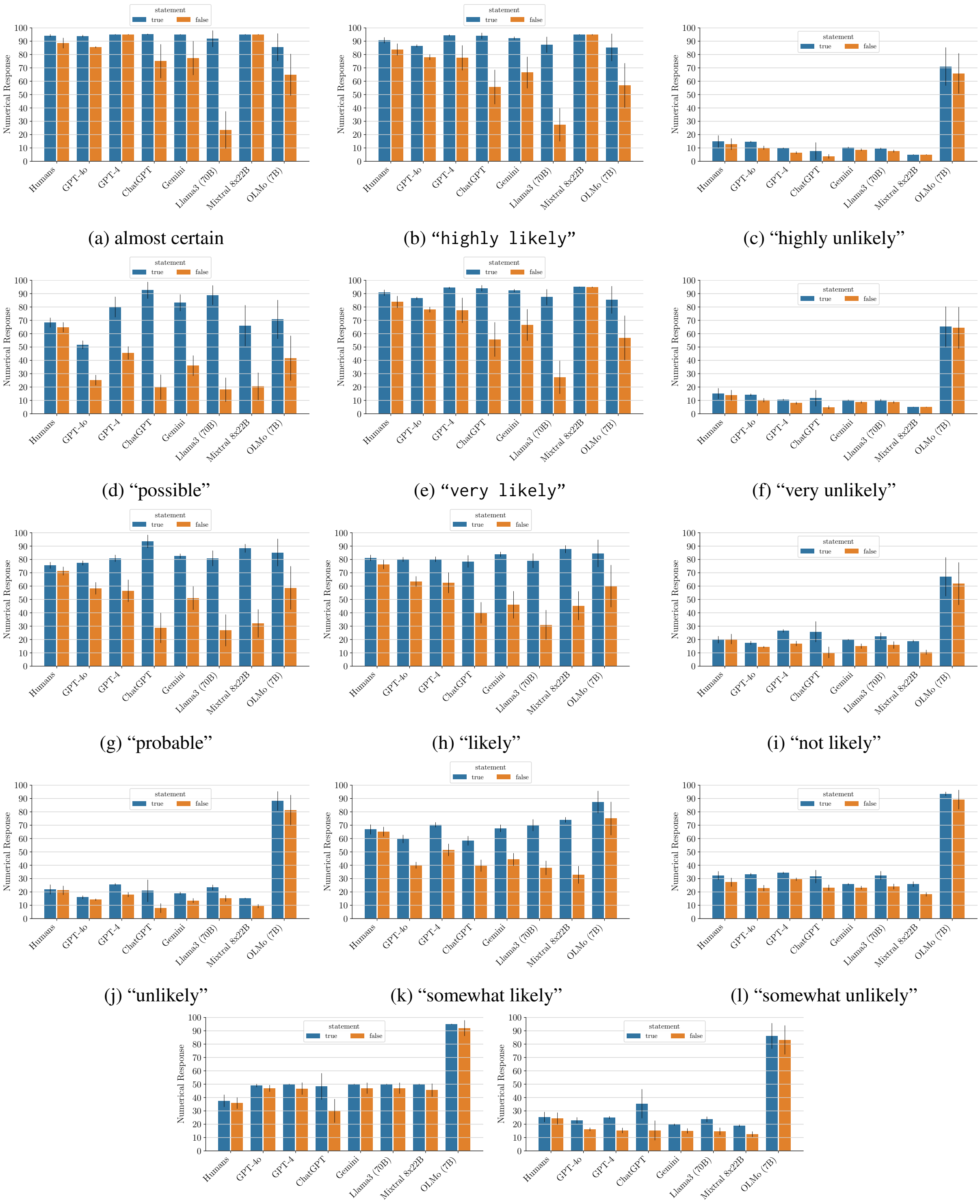


(k) `Mixtral 8x7B`



(l) `Mixtral 8x22B`

Figure 13: Empirical distributions of numerical probabilities per uncertainty expression in the non-verifiable setting. For each uncertainty expression (row), the empirical distribution is computed based on n=60 datapoints (for LLMs) and n=188 (for humans).



(a) almost certain (b) “highly likely” (c) “highly unlikely”

(d) “possible” (e) “very likely” (f) “very unlikely”

(g) “probable” (h) “likely” (i) “not likely”

(j) “unlikely” (k) “somewhat likely” (l) “somewhat unlikely”

(m) “uncertain” (n) “doubtful”

Figure 14: Mean response (and 95% confidence intervals) of verifiable statements discriminated by truthfulness of the statement across all 14 evaluated uncertainty expressions.

Table 8: Dissimilarities of model's empirical conditional distributions across verifiable settings. Lower results represent smaller distributional differences when comparing models' distribution.

	GPT-4o	GPT-4	ChatGPT	Gemini	Llama3 (70B)	Mixtral 8x7B	Mixtral 8x22B	OLMo (7B)	Gemma (2B)
Avg	1.9	1.8	9.1	4.2	8.8	7.6	3.7	13.1	48.5
almost certain	1.3	1.2	7.9	5.1	18.0	13.3	1.6	14.9	51.5
highly likely	1.5	1.3	12.8	5.1	17.1	11.2	2.2	18.9	55.1
very likely	1.6	1.8	10.6	7.0	16.0	13.3	1.8	18.4	53.6
likely	4.3	2.9	10.9	8.1	13.8	10.4	9.0	14.3	51.4
probable	3.2	3.2	17.2	5.8	14.4	16.0	10.6	12.8	50.7
somewhat likely	2.6	4.5	5.1	4.4	8.5	7.8	11.1	8.0	42.2
possible	4.8	3.4	14.6	7.5	11.6	11.2	6.6	15.0	48.6
uncertain	0.5	0.7	12.8	1.5	1.6	2.2	1.7	10.9	47.9
somewhat unlikely	2.3	0.4	3.6	2.0	2.1	4.2	1.6	14.0	34.2
unlikely	0.7	1.1	4.6	3.4	4.3	6.3	1.3	8.4	42.3
not likely	0.7	1.8	8.3	2.4	4.7	0.6	1.5	13.6	48.7
doubtful	0.2	1.5	14.6	2.5	5.1	6.6	1.9	24.7	47.1
very unlikely	1.2	0.8	2.1	1.4	2.6	1.1	0.2	3.1	50.7
highly unlikely	1.0	1.0	3.0	2.1	3.3	1.8	0.2	7.1	54.4

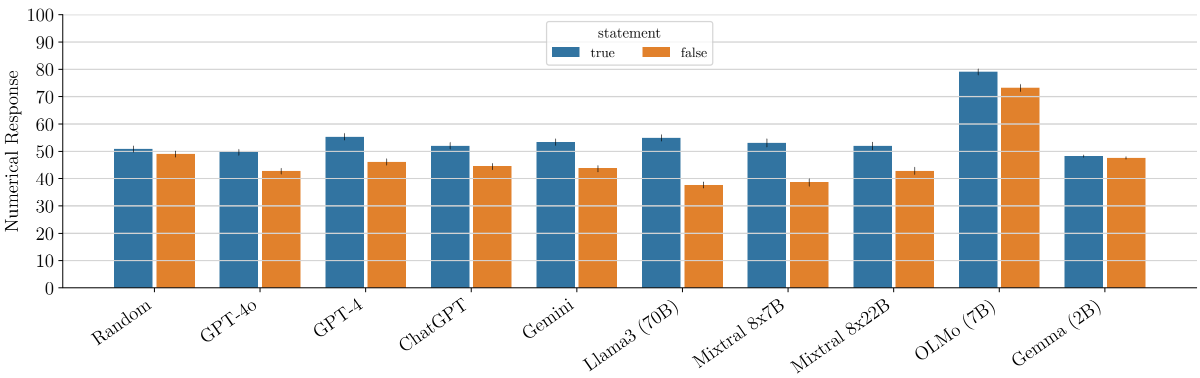
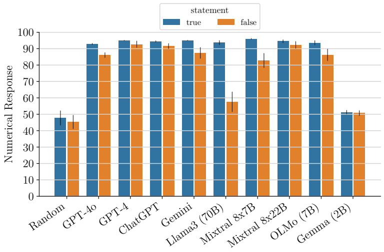


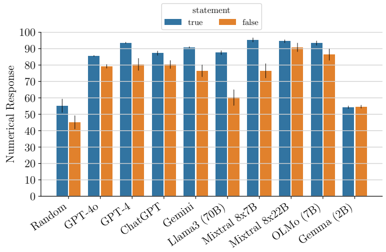
Figure 15: Mean response (and 95% confidence intervals) of verifiable statements across true and false statements averaged over the uncertainty expressions in the generalization set.

Table 9: Summary metrics average across uncertainty expressions using probabilistic decoding (`temperature=1`).

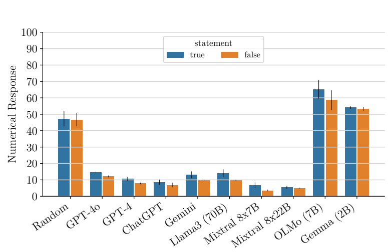
		Avg PA (↑)		Avg MAE (↓)		Avg Wasserstein-1 (↓)	
		NV	V	NV	V	NV	V
Human	Mode	27.6	27.6	—	—	—	—
	Individual	17.6	16.7	8.91	9.35	12.35	12.99
LLM	ChatGPT	16.4	12.8	6.40	8.32	7.65	12.14
	GPT4	24.4	21.4	4.62	4.00	9.78	6.72
	GPT4o	12.9	8.7	19.01	26.07	19.69	26.14



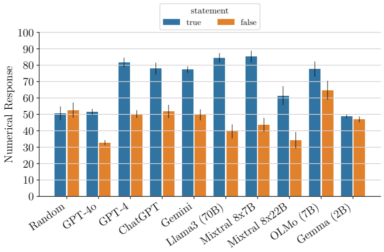
(a) almost certain



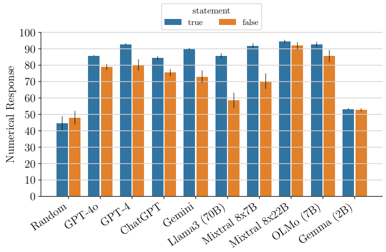
(b) “highly likely”



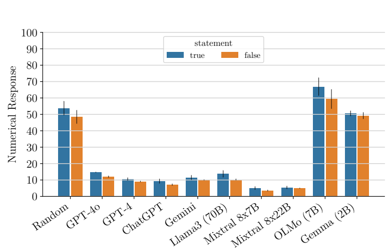
(c) “highly unlikely”



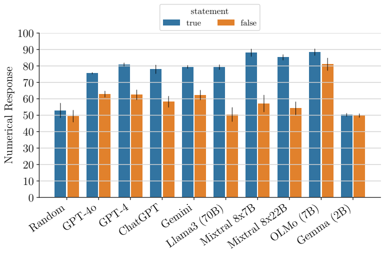
(d) “possible”



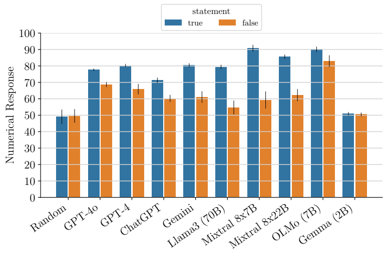
(e) “very likely”



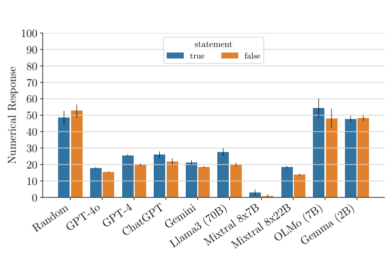
(f) “very unlikely”



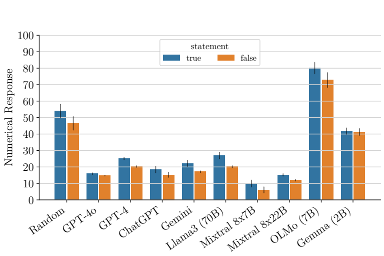
(g) “probable”



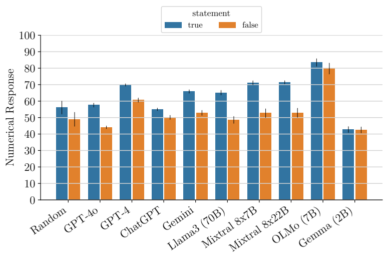
(h) “likely”



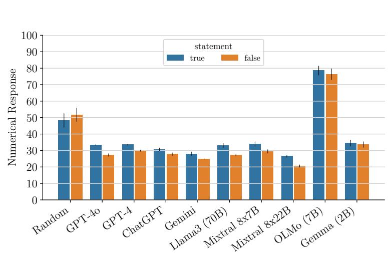
(i) “not likely”



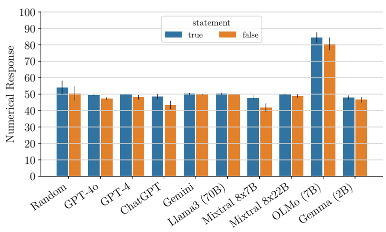
(j) “unlikely”



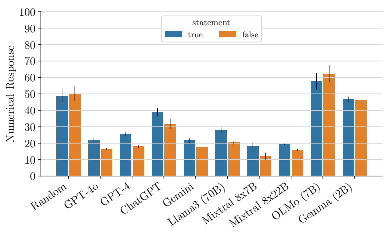
(k) “somewhat likely”



(l) “somewhat unlikely”

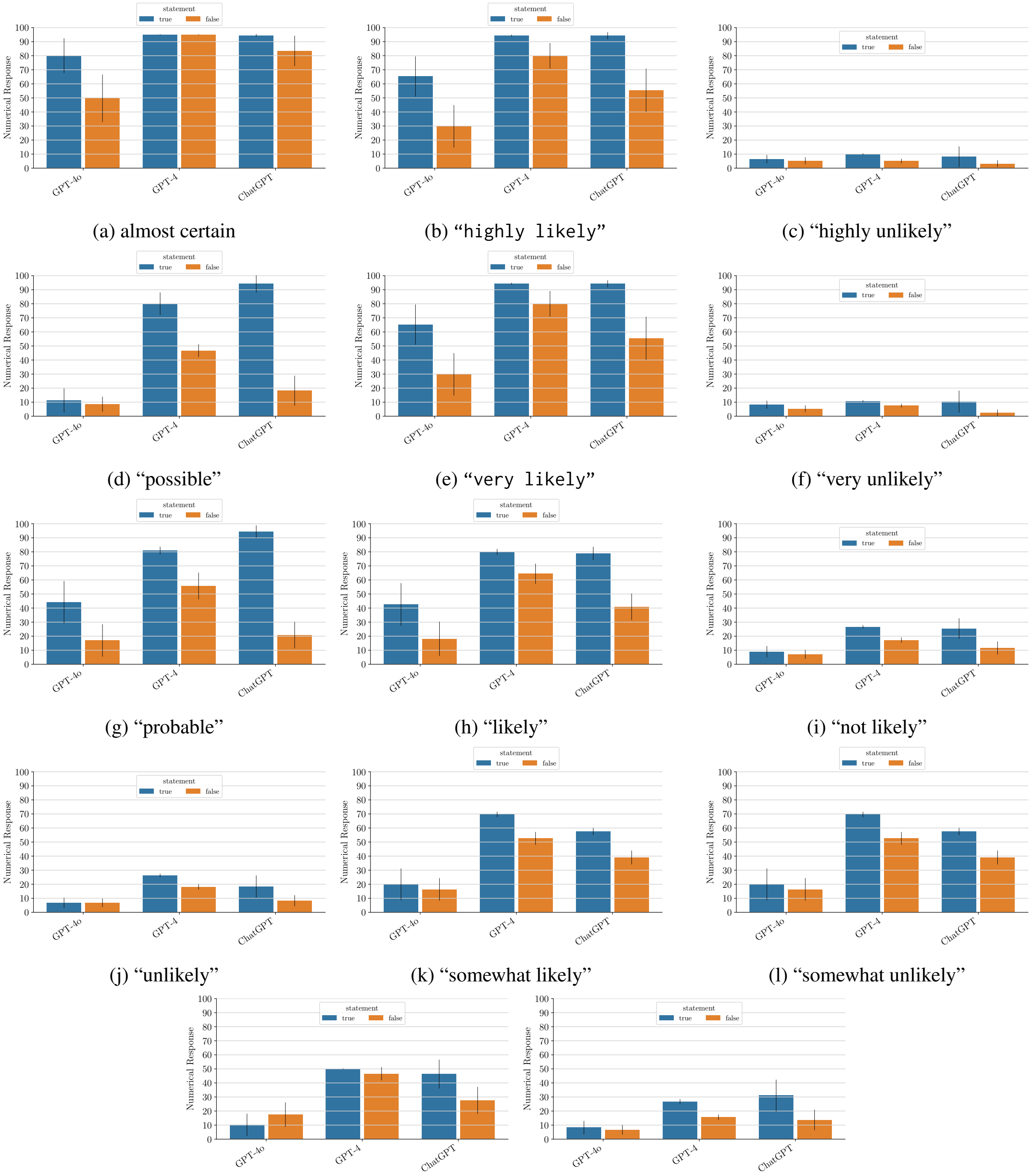


(m) “uncertain”



(n) “doubtful”

Figure 16: Mean response (and 95% confidence intervals) of verifiable statements across true and false statements for the 14 evaluated uncertainty expressions in the generalization set.



(a) almost certain

(b) “`highly likely`”

(c) “highly unlikely”

(d) “possible”

(e) “`very likely`”

(f) “very unlikely”

(g) “probable”

(h) “likely”

(i) “not likely”

(j) “unlikely”

(k) “somewhat likely”

(l) “somewhat unlikely”

(m) “uncertain”

(n) “doubtful”

Figure 17: Mean response (and 95% confidence intervals) of verifiable statements across true and false statements for the 14 evaluated uncertainty expressions, when using probabilistic decoding (i.e., `temperature=1`).