How Unlikely Is "Unlikely"? Assessing Verbal Probability Perception Across Large Language Models

Christos Petridis
christos.petridis@temple.edu
Department of Computer and Information Sciences
Temple University
Philadelphia, PA, USA

Konstantinos Pelechrinis
kpele@pitt.edu
Department of Informatics and Networked Systems
University of Pittsburgh
Pittsburgh, PA, USA

Zoran Obradovic
zoran.obradovic@temple.edu
Department of Computer and Information Sciences
Temple University
Philadelphia, PA, USA

ABSTRACT

Large language models increasingly produce and interpret verbal probability expressions, yet whether these expressions carry consistent meaning across models (or match human perceptions of uncertainty) remains unknown. We present a systematic cross-model evaluation using a word-to-number mapping task grounded in established human benchmarks. Eleven uncertainty expressions were presented to 19 models under two conditions, forced single-number response and explanation elicitation, alongside a novel bidirectional roundtrip test of internal consistency. LLMs track the human benchmark with surprising fidelity: word ordering is preserved, three anchor points are recovered, and "possible" shows the highest variance and cross-model disagreement of any expression tested, consistent with its documented bimodal interpretation in humans. However, models show a systematic upward bias for negative expressions such as "unlikely" and "improbable." Explanation elicitation reduces within-model variance while increasing between-model divergence, stabilizing individual models at the cost of inter-model consensus, and the roundtrip experiment reveals clear stratification, with frontier models maintaining coherent bidirectional representations. LLMs thus reproduce the structure of human verbal probability cognition, including its biases, while diverging systematically at the negative end—with implications for any setting where humans and models exchange probabilistic language.

1 INTRODUCTION

Every day, millions of people ask AI systems questions whose answers are uncertain:

- Will this treatment work?
- Is this investment risky?
- Is this news article trustworthy?
- Is this paper going to be accepted?

As large language models are quickly becoming collaborators in everyday decision making, people rely on them not only to retrieve information but also to evaluate uncertain situations, while weighing alternatives and making recommendations. In these interactions, communicating what is known is only part of the challenge, since equally important is communicating how certain that information is. In most cases AI systems communicate uncertainty using natural language ("likely", "unlikely","possible", "almost certain") rather than explicit probabilities [1, 6]. However, decades of decision science show that people interpret these expressions differently [2, 10–12]. Different people assign substantially different numerical interpretations to the same words, and those interpretations depend on context, expertise, and prior beliefs.

The question we want to answer with our work is whether the same is true for generative AI tools and how their interpretations compare to the human ones. Without shared interpretations, even perfectly accurate predictions may be misunderstood, leading users to overestimate or underestimate the confidence that an AI system intends to convey. Successful human-AI collaboration depends not only on AI making accurate predictions, but also communicating these predictions in a way that they are understandable to the users.

To investigate this, we conduct a systematic cross-model evaluation of verbal probability perception in large language models. We present 11 probability expressions to 19 models drawn from major commercial and open-source families, under two prompt conditions: a forced single-number response and an explanation elicitation condition. We ground our analysis in the human benchmark established by Mosteller and Youtz [4], which aggregates numerical interpretations of probability words across 20 human studies. We also extend our analysis with a novel bidirectional roundtrip experiment designed to test the internal consistency of each model's word-to-number mapping.

Our results indicate that LLMs collectively track the human benchmark with good fidelity. In particular, word ordering is preserved across all models and three anchor points (namely, "impossible", "even chance", and "certain") are recovered with near-zero variance. Furthermore, "possible" emerges as the single most variable expression across all models, a pattern consistent with the word's documented resistance of stable numerical interpretation in human studies. However, LLMs also rate probabilistic expressions as more probable than human norms suggest. The three anchor points above form the exception to this pattern. We also identify a larger, albeit marginal, inflation for negative expressions under the forced single number response condition. Explanation elicitation reduces within-model response variance but also increases divergence across models, while the roundtrip experiment reveals that while most models maintain a coherent bidirectional mapping between words and numbers, a small subset of them show mappings that are statistically indistinguishable from chance.

These findings have direct implications for human-AI communication. LLMs communicate uncertainty in ways that are, to a large extent, consistent with human norms, preserving word ordering, anchoring correctly at the extremes, and reproducing known patterns of ambiguity. Their deviations from those norms are real and predictable, concentrated in the non-anchor expressions examined, while the three anchor words (*impossible*, *even chance*, and

certain) show negligible bias. Within the examined expressions, deviation is largest for the inherently ambiguous *possible*, and we find weak evidence that negatively worded expressions exhibit slightly higher bias under the forced condition. Understanding where these deviations occur, and how they vary across models and prompt conditions, is a prerequisite for any application where probabilistic language is important.

2 RELATED WORK

The interpretation of verbal probability expressions has been studied for over half a century. Early work established that people assign widely varying numerical interpretations to the same words, and that those interpretations are stable across populations and contexts. Mosteller and Youtz [4] synthesized results from 20 studies covering 52 expressions, finding broad cross-population agreement for most words with one notable exception, namely, the word "possible". This word showed a distinctly bimodal distribution, with respondents splitting between near-zero and near-fifty interpretations. This instability has been consistently identified in subsequent work [3, 12]. Smithson et al. [5] formally documented a systematic asymmetry between positively and negatively worded expressions, showing that negative wording reduces precision in numerical translations independently of any shift in the mean response. Wintle et al. [12] replicated this pattern in a larger general-population sample, further showing lower consistency with normative guidelines for negative expressions, and attributed the effect to the greater difficulty of calibrating events that do not occur. Across studies, five expressions, namely, "very likely", "likely", "possible", "unlikely", and "very unlikely", show sufficiently overlapping numerical interpretations to form the basis of standardized verbal probability scales in clinical and intelligence communication [3]. However, "possible" is still a notable exception to this pattern, and, despite its consistent ordinal placement, it produces the widest distribution of numerical interpretations of any common probability expression, with respondents splitting between near-zero and near-fifty interpretations [4, 12].

More recent work examines how LLMs express and perceive uncertainty, spanning two related but distinct questions: how models interpret verbal probability language, and how reliably their expressed confidence tracks their actual accuracy or reasoning. Closest to our own work, Belem *et al.* [1] showed that LLMs map verbal probability expressions to numerical values in a broadly human-like manner, but that this mapping is systematically distorted by the model's prior beliefs about the associated statement. Specifically, they found that the same expression is assigned a higher probability when paired with a statement the model believes true than one it believes false. This bias is substantially larger than the corresponding effect observed in humans. A separate line of work examines not word-to-number mapping but the reliability of the models' expressed confidence more generally. Steyvers et al. [7] show that human users systematically overestimate the accuracy of LLM responses, and that longer explanations inflate users' perceived confidence independently of the model's actual accuracy. Steyvers and Peters [6] review this broader literature, identifying a persistent gap between models' implicit confidence, which is recoverable from token probabilities, and their explicit, verbalized confidence, with the latter being less reliable and more prone to overconfidence. Tanneru *et al.* [8] isolate one source of this gap in the context of chain-of-thought and token-importance explanations. The authors show that verbalized confidence scores are almost uniformly near-maximal regardless of whether the underlying answer is correct, making them uninformative, whereas an alternative probing-based measure correlates with both answer correctness and the faithfulness of the explanation. Our work is closest in spirit to [1] but differs in scope. Rather than testing whether a model's belief about a statement contaminates its report of a speaker's stated confidence, we isolate the lexical semantics of the probability words themselves, independent of speaker or statement context. This lets us compare LLM mappings directly against an established human benchmark, across a substantially larger and more current set of models, under both forced-response and explanation-elicitation conditions, and with a novel bidirectional consistency check.

3 EXPERIMENTAL SETUP

In this section we will describe in detail the different experimental settings we used in our study. We will also provide the reasoning behind these choices.

Stimuli: We use eleven verbal probability expressions spanning the full range of subjective certainty: impossible, very unlikely, improbable, unlikely, possible, even chance, probable, likely, very likely, almost certain, and certain. This set was chosen to follow the human benchmark established by Mosteller and Youtz [4], allowing us to directly compare to their aggregated meta-analytic norms.

Prompt Conditions: We evaluate each model under three prompting conditions. Each condition targets a different aspect of a model's verbal probability perception, the basic mapping itself, the effect of explicitly eliciting an explanation on that mapping, and the internal consistency of the mapping when queried bidirectionally.

(a) Basic forced-number prompt. Models are given each of the eleven words in turn and asked to output a single number between 0 and 100, with the endpoints anchored: 0 corresponds to an event that will never happen, and 100 to one that always happens. No explanation is requested.

Prompt 1

Estimate the probability that the event described by the word {word} WILL HAPPEN, on a scale from 0 to 100.
Scale definition:

- 0 = it will definitely NOT happen (0% chance of occurring)
- 50 = it is equally likely to happen or not (a coin flip)
- 100 = it will definitely happen (100% chance of occurring)

A HIGHER number always means MORE likely to happen. Respond with a single integer (the probability the event happens) and nothing else.

This condition establishes each model's baseline word-to-number mapping under minimal instruction, analogous to the forced-response paradigm used in the human decision-science literature. We also make the anchoring (0 and 100) explicit and unambiguous within

the instruction itself, rather than relying on the model to correctly infer scale direction. The reason for this, is that during some preliminary testing we observed that models sometimes invert the direction of the scale, treating higher numbers as indicating lower probability.

(b) Word-to-number mapping with explanation elicitation. This prompt is the same as the previous one, with the important addition of allowing the model to explain its choice step-by-step.

Prompt 2

Estimate the probability that the event described by the word {word} WILL HAPPEN, on a scale from 0 to 100.
Scale definition:

- 0 = it will definitely NOT happen (0% chance of occurring)
- 50 = it is equally likely to happen or not (a coin flip)
- 100 = it will definitely happen (100% chance of occurring)

A HIGHER number always means MORE likely to happen.
Respond with a JSON object with exactly two fields: "number" (an integer 0-100 = the probability the event happens) and “reason” (a short paragraph of 3-4 sentences). In your reasoning, always phrase the probability as the chance the event DOES happen, and make sure it is consistent with "number".
Respond with valid JSON and nothing else.

Comparing the plain and explanation-elicitation variants lets us test whether externalizing an explanation is associated with a change in the central tendency, the variance, or both, of a model’s response to a given word.

(c) Roundtrip (bidirectional consistency) prompt. The first two conditions test whether a model can go from word to number. The roundtrip condition tests whether that mapping is coherent in both directions. We first sample a number uniformly at random from [0, 100] and ask the model to name a probability word it associates with that number.

Prompt 3a

The number {number} represents a probability on a scale from 0 to 100, where 0 means “definitely will NOT happen” and 100 means “definitely WILL happen”. Give me a single word or short phrase (like “likely”, “almost certain”, “unlikely”) that best describes this probability.
Respond with only the word or phrase and nothing else.

In a separate, subsequent query, we present the model with the word it just generated and ask it to assign a number to that word.

Prompt 3b

What probability (0–100) does the word or phrase {word} represent, where 0 means “definitely will NOT happen”, 50 means “equally likely either way”, and 100 means “definitely WILL happen”?
Respond with a single integer and nothing else.

We define the **roundtrip error** as the absolute difference between the originally sampled number and the number recovered in this second query. A model with a perfectly coherent internal representation of probability language should recover a number close to the one it started from. In practice, a nonzero roundtrip error can arise from two distinct sources: a model may only ever express a small number of distinct probability values, in which case some error is unavoidable regardless of consistency, or the model’s word-to-number and number-to-word mappings may genuinely disagree with each other even when each direction looks reasonable in isolation. Rather than treating the roundtrip error as a single measure of inconsistency, we introduce baselines in Section 4.4 that separate these two sources.

Models: We evaluate the following models:

- **Anthropic:** Claude Haiku 4.5, Claude Sonnet 4.5, Claude Opus 4.5
- **OpenAI:** GPT-4.1, GPT-4o, GPT-4o-mini, GPT-5, GPT-5.1, GPT-5.2, GPT-5.4, GPT-5.4-mini, GPT-5.4-nano, GPT-5.5, GPT-OSS-20B
- **Google:** Gemma4 8B, Gemma4 26B
- **Meta:** Llama3-8B
- **Mistral AI:** Mistral-7B
- **Alibaba:** Qwen-14B

This set spans both proprietary and open-weight models, ranging from lightweight variants intended for low-latency deployment (e.g., GPT-5.4-nano, Gemma4 8B, Mistral-7B) to large frontier-scale models (e.g., GPT-5, Claude Opus 4.5), allowing us to assess whether verbal probability perception is a capability that scales with model size and access tier or one that is more uniformly present across the current model landscape.

For prompts 1 and 2 each one is sent 10 times to each model. For the roundtrip experiment we sample 30 numbers between 0 and 100 for each model. For each sampled number, we independently repeat the full roundtrip (Prompts 3a and 3b) five times, yielding 150 roundtrip trials per model in total. Repeating the full roundtrip, rather than only the second step, allows both the number-to-word and word-to-number directions to vary independently across repetitions of the same input. In all conditions, we do not set the temperature parameter explicitly, relying on each provider’s API default. We adopt this design deliberately. Most users interact with these models through consumer-facing chat interfaces that do not expose or modify the temperature parameter, so the default value reflects the conditions under which verbal probability expressions are actually produced and interpreted in practice. We note that default temperatures are not standardized across providers and may therefore differ across the models we evaluate.

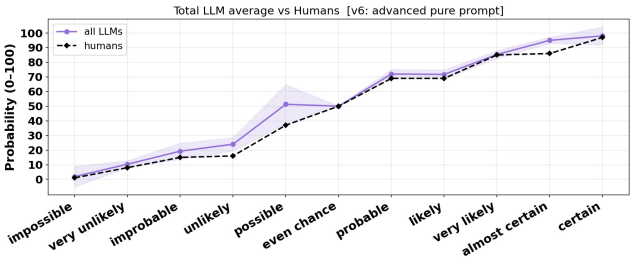


Figure 1: Aggregate probability scale across all evaluated LLMs versus the human benchmark, under prompt 1. Shaded regions denote ±1 standard deviation across models. The largest LLM–human divergence occurs at *possible*.

4 RESULTS

In this section we will present the results from our experiments and discuss their implications.

4.1 LLMs track the human benchmark closely

Figure 1 shows the aggregate mapping across all evaluated models, using Prompt 1, against the human benchmark established by Mosteller and Youtz [4]. The two curves are closely aligned across most of the scale. In particular, the word ordering is preserved, and the three anchor expressions, namely, *impossible*, *even chance*, and *certain*, are recovered with near-zero variance across models, closely matching the corresponding human anchors. The clearest departure from the human curve occurs at *possible*, where the aggregate LLM mean (≈51) sits well above the human mean (≈37) and carries the widest confidence band of any point on the curve. Despite the mismatch, this behavior is consistent with the bimodal and unstable interpretation of this word as documented in the human decision-science literature [4, 12], which we return to below.

We also examine whether the aggregate mapping obeys approximate numeric complementarity across antonym pairs, independent of any single word's absolute position on the scale. For a word w in our eleven-word set, let $\mu(w)$ denote the mean numerical response assigned to w, averaged across all models and trials under Prompt 1 (for the LLM curve) or across all human participants (for the human benchmark curve). For an antonym pair $(w, \bar{w})$, e.g., (*unlikely*, *likely*), we define the *complementarity gap* as:

$$D(w, \bar{w}) \;=\; \mu(w) + \mu(\bar{w}) - 100. \tag{1}$$

A pair with $D(w, \bar{w}) = 0$ is perfectly complementary, that is, the two antonyms partition the probability scale exactly in half, as would be the case if, e.g., $\mu(\text{unlikely}) = 20$ and $\mu(\text{likely}) = 80$. A nonzero value of $D(w, \bar{w})$ indicates that the pair's responses do not partition the probability scale into exact complements. This form of *subadditivity*, where judgments of complementary outcomes fail to sum to the full scale, aligns with a long-documented phenomenon in the judgment-under-uncertainty literature on complementary probability estimates, most notably the finding that explicitly unpacking an event into sub-events inflates its judged probability relative to the packed description [9]. While the mechanism in our setting (antonym word pairs rather than event unpacking) differs, the resulting failure of complementarity we will see in what follows is qualitatively similar. Because D is a property of the pair as a whole, it does not indicate which member of the pair, or in which direction, departs from exact complementarity.

Antonym pair $(w, \bar{w})$	D_{LLM}	SD across models	D_{human}
impossible / certain	−0.07	1.24	−2
very unlikely / very likely	−4.26	2.89	−7
improbable / probable	−8.75	6.91	−16
unlikely / likely	−4.29	5.37	−15

Table 1: Antonym complementarity gap $D(w, \bar{w}) = \mu(w) + \mu(\bar{w}) - 100$ for the 4 testable antonym pairs, using Prompt 1 across 19 models. Values closer to zero indicate greater complementarity, while negative values indicate sub-additivity. D_{LLM} is pooled across trials. (SD: standard deviation).

Of our eleven words, four pairs have a natural antonym partner: (*impossible*, *certain*), (*very unlikely*, *very likely*), (*improbable*, *probable*), and (*unlikely*, *likely*). The remaining three words do not have a clean antonym counterpart in our word list and are excluded from this analysis. This excludes the word *possible*, the word with the largest LLM–human divergence.

Table 1 reports $D(w, \bar{w})$ for both the aggregate LLM curve and the human benchmark. Both populations show sub-additivity across all four pairs. However, LLM pairs are consistently closer to perfect complementarity than the corresponding human pairs, most visibly for (*improbable*, *probable*) and (*unlikely*, *likely*).

4.2 Per-model variation & negative-word bias

Figures 2 and 3 break the aggregate pattern down by model, comparing the LLMs' responses to Prompt 1 against the same Prompt 2, in which models are additionally asked to articulate an explanation before their final answer (explanation elicitation). Across both conditions, most models track the human curve closely, but do show a consistent upward shift. For example, *unlikely* and *improbable* are assigned higher probabilities by most models than by human respondents, with `qwen_14b` and `llama3_8b` showing the largest deviations in both conditions. One model, `GPT-5.4 nano`, is a notable outlier at the *impossible* anchor, showing substantially wider variance under explanation elicitation than under Prompt 1 (we further examine this directly through its reasoning traces in Section 4.3).

Both negatively and positively worded expressions are assigned higher probabilities by LLMs than by human respondents (Wilcoxon signed-rank across 19 models, both $p < .01$ under both Prompts 1 and 2). However, while *likely* and *probable* are assigned higher values by LLMs than by humans, the gap margin is smaller than that of their negative counterparts (*unlikely*, *improbable*). Testing this asymmetry directly we find that under Prompt 1 the difference is on average about 2.1 percentage points over all 19 models (p-value for paired t-test 0.05), with a median of 0.9 (Wilcoxon signed-rank test $p = .067$). Under Prompt 2 the differences are not significant. Given the sample size for these tests we can clearly see that they are underpowered to detect small to moderate effects. We therefore treat this asymmetry itself as suggestive rather than established.

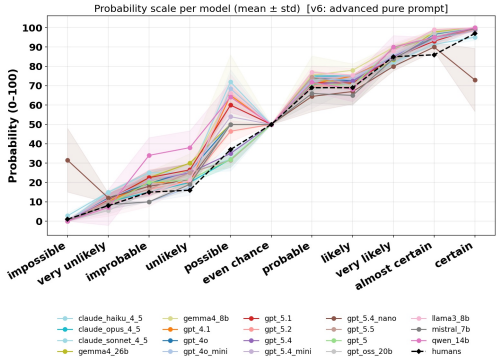


Figure 2: Probability scale per model (Prompt 1).

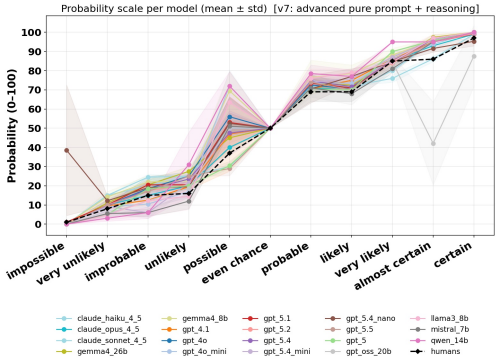


Figure 3: Probability scale per model (Prompt 2).

This is consistent with the antonym-complementarity result we saw earlier in Section 4 (Table 1). Both members of a pair shift upward by similar amounts, keeping LLM pairs closer to perfect complementarity than human pairs. However, as we can see from the standard deviation (SD) column in Table 1, there is substantial heterogeneity across individual models.

Comparing the results from the two prompts directly, we see that explanation elicitation visibly tightens the variance bands for most models, but it does not eliminate the upward shift. Furthermore, for a subset of models, mainly smaller open-weight models such as `qwen_14b` and `llama3_8b`, explanation elicitation appears to *increase* divergence from the human curve at *possible* and *probable* rather than reduce it. This suggests that explanation elicitation reduces the variance of each model's *individual* response distribution without necessarily pulling models toward consensus with either humans or each other. This instability effect seems to also be more pronounced in smaller, open-weight models than in frontier proprietary ones. While the polarity asymmetry we examined earlier is not robustly established, a paired comparison reveals a distinct and better-powered result. Each model's own gap is correlated across the two prompts ($r = .64$ across 19 models), so comparing a model against *itself* rather than against the population average removes much of the between-model noise that limits the two marginal tests. Under this paired comparison, the negative/positive gap shrinks under explanation elicitation for 13 of 19 models (Wilcoxon signed-rank $p = .006$). This indicates that whatever polarity asymmetry a given model exhibits, explanation elicitation tends to reduce it. This is a within-model claim that is statistically distinct from, and better supported than, the across-model, population-level question of whether the asymmetry exists at all.

4.3 Explanation elicitation does not uniformly reduce variance

Explanation elicitation (Prompt 2) tends to reduce response variance relative to Prompt 1. This effect is clearest for words such as *very unlikely* (Figure 4a), where every model with substantial baseline variance under Prompt 1 shows a clear reduction in variance under Prompt 2. The few cases in which variance appears to increase occur only among models whose variance is already negligible under both conditions. However there are exceptions, which are informative. There are two distinct patterns in which explanation elicitation is associated with *increased* variance, examined through qualitative inspection of the underlying reasoning traces.

Within-trial scale instability. For `GPT-5.4-nano` at *impossible*, variance increases substantially from Prompt 1 to Prompt 2 (Figure 4b). Reading the individual traces shows the model arriving at incompatible interpretations of the same instruction across trials. Some responses correctly reason toward a near-zero probability, others default to “`50%, given no additional context`” and at least one trial explicitly inverts the anchoring, reasoning that “`the prompt explicitly sets up the scenario to evaluate [...] I estimate a 90% chance`”. This is the scale-direction confusion our anchored prompt design is intended to prevent. However, here it re-emerges specifically *when the model is asked to explain itself*. This confusion appears inconsistently across trials rather than as a fixed per-model error. A related but distinct failure mode occurs for `gpt-oss-20b` at *certain* and *almost certain*. The model loses track of the target task itself when asked to explain its answer. In particular, in several trials this model evaluates the probability that the *prompt text itself* is naturally-occurring language, rather than the probability denoted by the target word. For instance, one trial's full reasoning reads:

> “`The phrase involves a nested quotation structure that would almost never appear in ordinary text unless typed deliberately or by mistake. Such an exact pattern is highly unlikely to occur naturally, making the event very rare. Consequently, the probability of this specific word sequence happening is only about one percent. Thus, it is considered quite improbable but not impossible.`”

Here, the model assigns a probability of 1, not because it judges *almost certain* to denote a low probability, but because it is evaluating how likely the *literal prompt text it received*, including its quotation marks, is to occur as a naturally-typed string. The target word itself never enters the model's explanation.

Latent disagreement surfaced by explanation elicitation. For several frontier models (e.g., `gpt_5.1`, `gpt_5.4` etc.) variance at *possible* increases sharply under explanation elicitation, even though these same models show near-zero variance at nearly every other word regardless of condition (Figure 4c). Unlike the previous pattern, this points to a principled rather than an arbitrary situation.

The word *possible* is documented in the human decision-science literature as eliciting an unusually unstable, bimodal split in interpretation [4, 12]. Asking these models to explain their answer may lead them to surface and act on multiple candidate interpretations across different trials, rather than collapsing to a single default response as they do under Prompt 1. We note that this is consistent with, but does not by itself establish, a directly parallel bimodal tendency in these models. Our data show increased *variance* across trials at this word, but we have not formally tested for multimodality (e.g., via a mixture-model fit) given our sample size per model.

Together, these patterns indicate that the variance-reducing effect of explanation elicitation is not uniform. Specifically, it depends on the word being evaluated, and on whether articulating an explanation supports genuine resolution of ambiguity (as with *possible*) or instead introduces new opportunities for the model to misconstrue the task or invert the intended scale.

4.4 Roundtrip consistency

We next performed the roundtrip experiment described in Section 3 (Prompts 3a and 3b). For each model, we sample 30 numbers uniformly from $[0, 100]$ and repeat the full roundtrip five times for each sampled number, yielding 150 roundtrip trials per model in total. We refer to the five repeated trials sharing a given sampled input as a *block*. Each model ends up with 26 blocks each, instead of 30 blocks, since we sample integers in this range and our sampling yields 26 distinct sampled values.

For each model m, let $x_i \sim \text{Uniform}[0, 100]$ be the i-th sampled probability presented in the first roundtrip query (3a), let $w_i = f_m(x_i)$ denote the word returned by model m for that value, and let $\hat{x}_i = g_m(w_i)$ denote the number model m assigns to that same word w_i when queried independently in the second roundtrip step (3b). We define the roundtrip error for trial i as:

$$e_i = |x_i - \hat{x}_i|, \tag{2}$$

and the *mean absolute roundtrip error* for model m as:

$$\text{MARE}_m = \frac{1}{N} \sum_{i=1}^{N} e_i, \tag{3}$$

where $N = 150$. A model with a perfectly self-consistent word–number mapping would achieve $\text{MARE}_m = 0$.

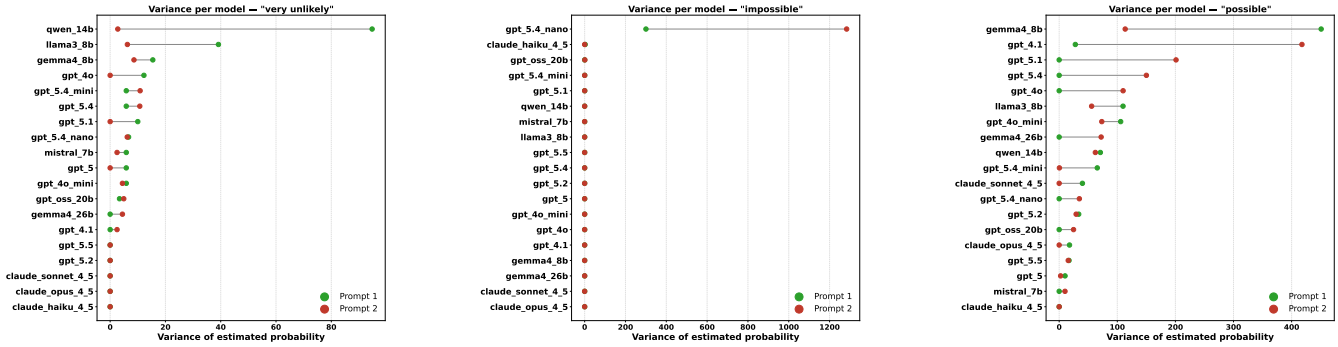
Baselines. A natural first reference point is a naive independence baseline, obtained by treating both x_i and $\hat{x}_i$ as independent draws from $\text{Uniform}[0, 100]$. This yields an expected error that is equal to the expected absolute difference of two independent uniform random variables, that is, $100/3 \approx 33.3$. However, this baseline is not meaningful in practice. It implicitly assumes the recovered value $\hat{x}_i$ is itself uniformly distributed, which no model in our data satisfies. A model that simply ignores the input and always answers "equally likely" (recovering 50 every time) achieves $\mathbb{E}|X - 50| = 25$ for $X \sim \text{Uniform}[0, 100]$, thus, beating the naive baseline of 33.3 while carrying absolutely no information about the input. We therefore report two corrected reference points instead: (a) the *best-constant baseline* of 25, and, (b) a *quantization floor* of $25/k$ for a model that only ever produces k distinct recovered values, representing the unavoidable error from having a limited vocabulary even if that vocabulary is used with perfect internal consistency (the derivation of the quantization floor is in Appendix A). Any excess of a model's observed MARE_m above its own quantization floor reflects genuine inconsistency in *how* it uses its available vocabulary, not the size of that vocabulary itself. We also report a permutation-based test of whether each model's observed MARE is distinguishable from chance. In particular, we shuffle the pairing between each model's own sampled inputs and recovered values 5,000 times, preserving the model's actual empirical output distribution, and report the fraction of shuffles that achieve a MARE at least as low as the model's observed value. A large fraction here indicates that a model's roundtrip mapping is statistically indistinguishable from chance, given only its own tendency to produce certain output values regardless of input.

Model	k	Floor	MARE	Gap	p
gpt_5.5	15	1.67	4.49	2.82	<.0001
gpt_5	17	1.47	4.63	3.16	<.0001
gpt_5.4	9	2.78	5.02	2.24	<.0001
gpt_5.1	19	1.32	5.19	3.87	<.0001
claude_sonnet_4_5	9	2.78	6.02	3.24	<.0001
claude_opus_4_5	15	1.67	6.63	4.97	<.0001
gpt_4.1	14	1.79	6.74	4.95	<.0001
gpt_4o	14	1.79	7.05	5.26	<.0001
gpt_5.2	13	1.92	7.15	5.23	<.0001
gemma4_26b	17	1.47	7.61	6.14	<.0001
gpt_oss_20b	23	1.09	8.27	7.18	<.0001
gpt_5.4_mini	15	1.67	8.98	7.31	<.0001
gpt_4o_mini	12	2.08	12.81	10.72	<.0001
claude_haiku_4_5	14	1.79	15.02	13.23	<.0001
gemma4_8b	20	1.25	20.69	19.44	<.0001
qwen_14b	8	3.12	26.31	23.19	.0001
gpt_5.4_nano	10	2.50	26.77	24.27	<.0001
mistral_7b	8	3.12	27.01	23.89	**.1024**
llama3_8b	12	2.08	27.71	25.63	**.0090**†

Table 2: Roundtrip consistency results across all 19 models, ranked by observed MARE. p is the permutation-test p-value (5K resamples). Bold values do not survive Bonferroni correction across the 19 models tested; † indicates a value that survives the less conservative Holm-Bonferroni correction.

Results. Table 2 reports, for each model, its empirical vocabulary size k (the number of distinct recovered values observed across all 150 trials), the resulting quantization floor, the observed MARE, the gap between the two, and the permutation test result.

Two patterns are visible that a raw MARE ranking alone would obscure. First, only four models, namely, `GPT-5.4 nano`, `Qwen-14b`, `Mistral-7B`, and `llama3-8b`, fail to beat the constant baseline of 25. Also, `Gemma4-8B`, despite the largest gap in the table, still achieves a lower MARE than a strategy that ignores the input entirely. Second, the permutation test separates these four worst-performing models from one another despite their similar raw MARE values. Given that we run this test across 19 models, we apply a Bonferroni



(a) ***very unlikely*** **(b)** ***impossible*** **(c)** ***possible***

Figure 4: Variance of estimated probability per model, comparing Prompt 1 (no explanation, green) and Prompt 2 (explanation elicitation, red). (a) ***very unlikely*** **shows the general stabilization pattern. Every model with substantial baseline variance shows a clear reduction under explanation elicitation. (b)** `gpt_5.4_nano` **shows a sharp variance increase at** ***impossible,*** **driven by within-trial scale instability. (c) Several frontier models show increased variance at** ***possible*** **despite near-zero variance elsewhere, consistent with explanation elicitation surfacing latent disagreement documented for this word in humans.**

correction ($\alpha/19 \approx .0026$) to guard against false positives from multiple testing. Under this correction, `Qwen-14b` and `GPT-5.4 nano` remain clearly distinguishable from chance. Essentially, whatever these two models are doing is not fully random, even though they perform worse than ignoring the input. Connecting to our earlier results, `GPT-5.4 nano` is the same model whose reasoning traces we examined in Section 4.3 for scale-direction instability at *impossible*. Its roundtrip failure here is consistent with that finding, though driven specifically by its word-to-number mapping (Step 3b) rather than its choice of word (Step 3a), where as we discuss in what follows its distinct-word ratio of .30 is only mid-pack (see Table 3). `llama3-8b` performance is borderline, as it survives a sequential Holm-Bonferroni correction ($p = .009$) but not the stricter flat Bonferroni threshold, so we treat its apparent coupling as suggestive rather than conclusive. Finally, `Mistral-7B`'s mapping is not distinguishable from chance under any correction ($p = .102$). That is, we cannot statistically distinguish this model's roundtrip mapping from one that ignores the input entirely.

We further measure how many distinct words a model returns in Step 3a across the five trials within a block, normalized by the number of trials in that block to account for the repeated sampling of 4 out of 30 initial integers. This diagnostic targets a different property than the empirical vocabulary size k used in our quantization-floor analysis. Whereas k counts distinct *final* recovered numbers pooled across all trials, the word-choice diversity ratio is computed *within* a block of trials sharing an identical input, so any variation it detects reflects pure inconsistency rather than legitimate resolution across distinct inputs. As we can see in Table 3, `GPT-5.4`, `GPT-4.1` and `Claude-Opus-4.5` pick nearly the same word every time (a mean distinct-word ratio of .20, .21 and .22, respectively), while `Gemma4-8B` shows the least stable word choice of any model in our results (.59). Despite this, `Gemma4-8B` still beats the constant baseline overall, indicating that its number-mapping step (3b) compensates for highly unstable word choice.

Within-family comparisons. Focusing on models from the same family allows us to isolate the effect of scale from confounds of architecture and training pipeline. Of course, this isolation is not perfect, since same-family variants can still differ in fine-tuning data and other choices beyond parameter count, but nonetheless is informative. Within the Gemma4 family, the 26B variant achieves a substantially lower roundtrip error than the 8B variant (7.61 vs. 20.69). A similar, but less pronounced, pattern holds within the GPT-4o family, where the full-size model outperforms its mini variant (7.05 vs. 12.81). Within the GPT-5 family, we restrict this comparison to the three variants whose naming denotes an explicit size tier (`gpt_5`, `gpt_5.4_mini`, `gpt_5.4_nano`). Roundtrip error increases

Model	Mean distinct-word ratio
gpt_5.4	.196
gpt_4.1	.208
claude_opus_4_5	.219
gpt_5.2	.242
gpt_4o_mini	.246
gemma4_26b	.250
gpt_4o	.250
claude_sonnet_4_5	.258
gpt_5.1	.273
gpt_5.4_mini	.277
gpt_5.5	.277
claude_haiku_4_5	.285
gpt_5	.285
gpt_5.4_nano	.300
qwen_14b	.350
mistral_7b	.354
gpt_oss_20b	.427
llama3_8b	.438
gemma4_8b	.588

Table 3: Mean distinct-word ratio per block (five trials sharing an identical sampled input, normalized by block size), ranked from most to least stable, across all 19 models.

monotonically across the three size tiers: 4.63, 8.98, 26.77. Even the same-tier point releases (`gpt_5.1`, `gpt_5.2`, `gpt_5.4`, `gpt_5.5`), all cluster tightly between 4.49 and 7.15 regardless of version number. These are all lower than `gpt_5.4_mini` and `gpt_5.4_nano`, and consistent with these point releases being updates within the same size tier rather than different model sizes. The three Claude variants follow a similar, though not perfectly monotonic, ordering by size (`claude_opus_4_5`: 6.63, `claude_sonnet_4_5`: 6.02, `claude_haiku_4_5`: 15.02).

These comparisons provide more direct evidence for an effect of scale on roundtrip consistency than cross-family comparisons alone, where architecture and training differences could otherwise account for the observed gap. We note, however, that with only a few models per family, we cannot distinguish a robust scale effect from incidental variation, and a larger sample of same-family model sizes would be needed to establish this more rigorously.

5 DISCUSSION AND LIMITATIONS

Taken together, our results point to a consistent picture, where LLMs inherit the "structure" of human verbal probability perception without fully inheriting its calibration. More specifically, word ordering and anchor fidelity are reproduced across nearly every model we evaluate, regardless of scale or provider, and the highest variance and cross-model disagreement in our results occur specifically at *possible*, which is consistent with the bimodal instability documented for this word in humans. This suggests that this structure is a robust and learned property of the distribution of probability language in pretraining data rather than a capability that requires scale or explicit alignment. At the same time, models diverge from humans in a specific and repeatable way. LLMs systematically rate both negatively and positively worded expressions as more probable than human norms suggest, with a marginal trend toward this inflation being larger for negative expressions specifically. While explanation elicitation improves each model's *internal* consistency (narrower response variance), it does not correct this bias and, for a subset of models, appears to increase divergence from the human curve rather than reduce it. This might serve as an indication that eliciting explanation stabilizes a model's individual mapping without necessarily correcting it or moving models toward consensus with one another. This connects to a broader concern raised in prior work on LLM reasoning outside the probability domain. In particular, chain-of-thought prompting has been found to be able to induce behavioral shifts, such as collapsing onto a narrow set of answer patterns, that are difficult to distinguish from genuine improvements in task understanding [13]. Our transcript-level evidence in Section 4.3 suggests a similar risk applies to reasoning about verbal probability expressions. For at least two models, the instructions to provide an explanation appear to introduce new failure modes (task misidentification and within-trial scale inversion) rather than resolving genuine uncertainty about the target word.

The roundtrip experiment aims at quantifying the self-consistency of each model, i.e., whether a model's word-to-number and number-to-word mappings agree with each other, rather than at tracking calibration accuracy (i.e., how closely those mappings match human values). Most models maintain a coherent bidirectional representation of probability language, while a small subset, spanning both proprietary and open-weight families, show mappings that are statistically indistinguishable from chance. Within-family comparisons (e.g., Gemma4 26B vs. 8B, GPT-4o vs. GPT-4o-mini, and a three-tier ladder within the GPT-5 family) suggest that this self-consistency gap is at least partly attributable to scale, independent of architecture or training pipeline.

Human-AI communication: The positive bias documented above, and specifically the potentially larger inflation for negatively worded expressions, has a direct practical consequence. This bias operates on the grammatical polarity of the expression itself, independent of whether the event being described is desirable or undesirable. A human reader will tend to *underestimate* the probability a model actually assigns whenever it hedges using negative wording, regardless of whether that wording describes an undesirable event ("side effects are *unlikely*"), in which case the reader is given false reassurance, or a desirable one ("recovery is *unlikely*"), in which case the reader is led to be more pessimistic than the model's own estimate warrants. Positive-worded hedges (*likely*, *probable*) show the same directional bias but to a smaller degree, so the risk of underestimation is reduced, not eliminated, when a model hedges using positive language. In domains where such hedges carry real consequences, this systematic inflation is very important. Since a model's own mapping for a hedge word sits above the corresponding human mapping, this can lead to the probability underestimation mentioned above. Whether this is false reassurance or undersold good news depends on the desirability of the event being described. A clinical assistant describing side effects as *unlikely*, or a safety system reporting a fault as *improbable*, risks leaving a patient or operator more confident in a good outcome than the model's own estimate warrants. In the opposite direction, a financial tool describing a recovery as *likely*, or a forecasting system describing an opportunity as *probable*, risks understating the model's own confidence in a desirable outcome, leading a user to discount good news more than the model's estimate would justify. This is a distinct mechanism from the reliability-overestimation effects documented elsewhere in the literature, where users trust an LLM's correctness more than its actual accuracy warrants, often driven by superficial cues such as response length [7]. The bias we report does not concern whether a model's underlying claim is accurate, but whether the numeric belief a model intends to convey through a hedge word is the same as that of the human user. Even a model with perfectly calibrated internal probabilities could, through this word-choice mismatch alone, leave a human reader with a systematically different estimate than the model's own internal state would justify.

Limitations: As aforementioned, we query each model at its default temperature setting in order to reflect how most users interact with these systems through consumer-facing interfaces. However, default temperatures are not standardized across providers, so some of the cross-model variance we report may be partly confounded with each model's baseline stochasticity rather than reflecting a pure effect of reasoning or model tier. Additionally, each triplet (model, word, condition) is estimated from only $N = 10$ samples. While we believe that this is sufficient to establish the directional patterns we report, individual per-word means, especially for the smaller open-weight models with wider variance, should be interpreted with appropriate uncertainty. Finally, our forced-response

and explanation elicitation prompts use a single fixed instruction template. We have not tested robustness to paraphrasing the anchoring instruction itself, so we cannot rule out that some fraction of the observed bias is sensitive to the specific wording we chose rather than being a fully prompt-invariant property of these models.

Our human benchmark is drawn from the meta-analytic norms of Mosteller and Youtz [4], aggregated across studies conducted well before the emergence of current LLMs. Since we rely on this existing benchmark, some of the divergence we attribute to LLMs could in principle reflect true shifts in population-level human interpretation since these norms were established, rather than a property specific to LLMs. Finally, our findings are necessarily specific to the eleven expressions, three prompt conditions, and the models we evaluate. Extending this analysis to a broader vocabulary of uncertainty language, additional prompt phrasings, and the continually evolving landscape of available models is an important direction for future work. A further extension worth pursuing is *comparative* and *contextual* probability language, which convey a probability *relative* to some reference point rather than an absolute value. This includes expressions such as *more likely than*, *slightly more probable*, or *far less likely*. The same comparative phrase can correspond to very different absolute magnitudes depending on context (e.g., "more likely" could indicate a shift from 10% to 15%, or from 40% to 90%). This is a form of ambiguity entirely distinct from the context-free lexical mapping we study here, and one that may interact with the biases we report in ways that context-free elicitation cannot reveal.

Finally, two further metrics would strengthen the results but require substantially more data than we collected in this study. First, an entropy-based measure of word-choice consistency would be more informative than the distinct-word ratio, since it could distinguish a model that splits its answers at a ratio of 4-to-1 between two words from one that splits them evenly. Second, a mutual information measure between the sampled input and the word or number returned would directly quantify how much information survives the roundtrip, incorporating much of what k, the quantization floor and the permutation test currently do separately into a single statistic. Both metrics would be feasible with data on the order of thousands, rather than hundreds, of roundtrip trials per model, which we identify as a concrete direction for future work.

6 CONCLUSIONS

LLMs largely inherit the structure of human verbal probability perception (word ordering, anchor fidelity, a rich vocabulary in frontier models) but not its calibration. Both negatively and positively worded expressions are overrated relative to human norms, with only a marginal, non-robust trend toward this being larger for negative expressions specifically. Also, neither explanation elicitation nor scale reliably fixes the underlying inflation. Our experiments also indicate that self-consistency and calibration accuracy are separate, only loosely related properties. These results suggest that current LLMs communicate uncertainty in a recognizably human-like way on the surface, but carry specific, measurable distortions that matter wherever hedged language is trusted at face value.

REFERENCES

[1] Catarina G Belem, Markelle Kelly, Mark Steyvers, Sameer Singh, and Padhraic Smyth. 2024. Perceptions of linguistic uncertainty by language models and humans. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Miami, Florida, USA, 8467–8502.

[2] David V Budescu and Thomas S Wallsten. 1985. Consistency in interpretation of probabilistic phrases. *Organizational behavior and human decision processes* 36, 3 (1985), 391–405.

[3] M Jawad Hashim. 2024. Verbal probability terms for communicating clinical risk-a systematic review. *The Ulster Medical Journal* 93, 1 (2024), 18.

[4] Frederick Mosteller and Cleo Youtz. 1990. Quantifying probabilistic expressions. *Statist. Sci.* 5, 1 (1990), 2–12.

[5] Michael Smithson, David V Budescu, Stephen B Broomell, and Han-Hui Por. 2012. Never say "not": Impact of negative wording in probability phrases on imprecise probability judgments. *International journal of approximate reasoning* 53, 8 (2012), 1262–1270.

[6] Mark Steyvers and Megan AK Peters. 2026. Metacognition and uncertainty communication in humans and large language models. *Current Directions in Psychological Science* 35, 3 (2026), 131–139.

[7] Mark Steyvers, Heliodoro Tejeda, Aakriti Kumar, Catarina Belem, Sheer Karny, Xinyue Hu, Lukas W Mayer, and Padhraic Smyth. 2025. What large language models know and what people think they know. *Nature Machine Intelligence* 7, 2 (2025), 221–231.

[8] Sree Harsha Tanneru, Chirag Agarwal, and Himabindu Lakkaraju. 2024. Quantifying uncertainty in natural language explanations of large language models. In *International Conference on Artificial Intelligence and Statistics*. PMLR, PMLR, Valencia, Spain, 1072–1080.

[9] Amos Tversky and Derek J Koehler. 1994. Support theory: A nonextensional representation of subjective probability. *Psychological review* 101, 4 (1994), 547.

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

[11] 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 (1993), 135–138.

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

[13] Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. 2023. Causal parrots: Large language models may talk causality but are not causal. arXiv:2308.13067. arXiv:2308.13067 [cs.AI]

A QUANTIZATION FLOOR DERIVATION

During the roundtrip experiment we noted that some models recover only a small number of distinct values across the roundtrip procedure, regardless of the original sampled input. For example, a model might only ever express its answer using a handful of familiar round-number anchors (e.g., 25, 50, 75, and 95, say) rather than the full continuum between 0 and 100. Such a model cannot achieve zero roundtrip error even if it is otherwise perfectly consistent. For example, if the original input was 43, and the model always rounds to the nearest of its four available anchors (50), it will report 50 regardless of how many times the roundtrip is repeated. This is not evidence of an incoherent mapping, but rather it is an unavoidable consequence of having only a few values *available* to report. We derive here the smallest error such a restriction can produce, so that a model's excess error above this bound can be attributed to genuine inconsistency rather than vocabulary size.

Formally, a model that only ever reports one of k distinct values is equivalent, for this purpose, to a *quantizer*: a rule that maps a continuous input to the nearest of k discrete representative values. Suppose such a quantizer partitions $[0, 100]$ into k equal-width bins, each of width $w = 100/k$, and reports the midpoint of whichever bin the true input falls into. If the input $X \sim \text{Uniform}[0, 100]$, then within any one bin, X is uniformly distributed on an interval of width w centered at that bin's midpoint. Writing X relative to the midpoint as $X' \sim \text{Uniform}[-w/2, w/2]$, the mean absolute error contributed by this bin is:

$$\mathbb{E}|X'| = \frac{1}{w}\int_{-w/2}^{w/2} |x|\,dx = \frac{1}{w}\left[2\int_{0}^{w/2} x\,dx\right] = \frac{1}{w}\cdot\frac{w^2}{4} = \frac{w}{4}. \quad (4)$$

By symmetry, this is the same for every bin, so it is also the overall mean absolute error across the full range. Substituting $w = 100/k$:

$$\text{floor}(k) = \frac{100/k}{4} = \frac{25}{k}. \quad (5)$$

This is the error a model would incur *purely from having only k distinct values available*, even if it places those k values optimally and reports them with perfect consistency. It therefore serves as a lower bound, not a prediction. A model's observed MARE_m can equal $\text{floor}(k)$ only if its limited vocabulary is used as efficiently as possible. The quantization floor represents the best theoretically achievable MARE for a k-level representation under a uniform input distribution. Excess error can result from either non-optimal placement/use of representational anchors or genuine mapping inconsistency.