Beyond 0–100: How Confidence Scale Design Shapes Verbalized Confidence and LLM Metacognition

Yuyang Dai
INSAIT
y9657422@gmail.com

Yuxia Wang
INSAIT
yuxia.wang@insait.ai

Abstract

Verbalized confidence, in which LLMs report a numerical certainty score, is widely used for uncertainty estimation and calibration in black-box settings, yet the confidence scale itself (typically 0–100) is rarely examined. We show that this design choice is not neutral. Across six primary LLMs and three datasets, with an additional dense-architecture model used for robustness analysis, we find that verbalized confidence is heavily discretized: the three most frequent values account for 78.2%–92.1% of responses under the standard scale. To investigate this phenomenon, we manipulate confidence scales along three dimensions: granularity, boundary placement, and range regularity, and evaluate metacognitive sensitivity using $meta\text{-}d'$. Within the evaluated models, datasets, prompts, and decoding conditions, a 0–20 scale improves metacognitive efficiency over the standard 0–100 format, while aggressive boundary compression degrades performance and round-number preferences persist even under irregular ranges. These results show that confidence scale design affects the quality of verbalized uncertainty in our experiments and should be treated as a first-class experimental variable in LLM evaluation.

1 Introduction

As Large Language Models (LLMs) are increasingly integrated into decision-making pipelines, reliably estimating their confidence or uncertainty becomes critical. In black-box settings, ***verbalized confidence***, where models report a numerical certainty score via prompting, has become the predominant paradigm (Xiong et al., 2024): confidence scores are used to decide when to abstain from answering in selective prediction (Kadavath et al., 2022), to trigger self-correction on low-confidence outputs (Li et al., 2024), and to calibrate how much end users should trust a given response (Zhou et al., 2024). Yet the reliability of these confidence signals remains poorly understood.

Given its ubiquity, research on verbalized confidence has focused on elicitation strategies such as Chain-of-Thought prompting (Wei et al., 2022), self-consistency (Wang et al., 2022), and training models to express uncertainty (Lin et al., 2022a), while treating the ***confidence scale*** itself (typically 0–100) as a neutral instrument. Yet psychometric research has long established that scale granularity and anchoring alter the quality of self-assessment (Lozano et al., 2008; Preston and Colman, 2000). Whether analogous effects hold for LLM judges remains unexplored.

We find that LLM confidence reports exhibit a striking phenomenon we term *confidence discretization*: LLMs do not use the 0–100 scale as a continuous spectrum but cluster outputs around a small set of round-number anchors. This compression, which we find to be pervasive yet model-specific, suggests that verbalized confidence may be shaped as much by token-level biases (Zhao et al., 2021; Shaki et al., 2023) as by genuine self-evaluation, potentially distorting standard calibration metrics such as ECE (Guo et al., 2017). This raises a central question: *does confidence scale design affect the quality of verbalized uncertainty signals in LLMs?*

To investigate, we present the first systematic empirical study of confidence scale design for LLM metacognition, manipulating scales along three dimensions: *granularity* (0–5 to 0–100), *boundary shifting* (e.g., lower bound raised to 20 or 60 (Tversky and Kahneman, 1974)), and *semantic robustness* via non-standard ranges (e.g., 0–73, 14–86 (Pellert et al., 2024)). We adopt $meta\text{-}d'$ from signal detection theory (Maniscalco and Lau, 2012; Fleming and Lau, 2014) to quantify metacognitive sensitivity independent of overall bias. Our contributions are:

- We reveal the vulnerability of verbalized confidence to scale design, identifying **confidence**

discretization, a robust phenomenon across multiple LLM families.

- Confidence scale granularity significantly modulates metacognitive quality, identifying a consistent sweet spot 0–20 that outperforms the standard 0–100 scale.
- Through non-standard scales, we show that LLMs exhibit limited adaptation to numerical range boundaries (*e.g., raising the lower bound to 60 as* $[60, 100]$ *leaves models clustered near the ceiling*), providing practical guidelines for confidence scale design.

2 Related Work

Verbalized Confidence and Calibration As LLMs became accessible primarily through black-box APIs, uncertainty estimation shifted from logit-based methods (Kadavath et al., 2022) to verbalized confidence, where models report certainty as a numerical score (Lin et al., 2022a; Xiong et al., 2024). Prior work has focused on improving elicitation, via Chain-of-Thought prompting (Wei et al., 2022), metacognitive prompting (Wang and Zhao, 2024), self-consistency (Wang et al., 2022), and RLHF-specific strategies (Tian et al., 2023), and on evaluation via Expected Calibration Error (ECE: Guo et al., 2017). This literature treats the confidence scale itself as a fixed, neutral backdrop. However, we find that LLMs do not use the 0–100 scale as a continuous spectrum but cluster outputs around a small set of round-number anchors. Then, when models concentrate the vast majority of their reports on two or three values, ECE bins become sparse or empty across most of the range, making calibration estimates unreliable. This *discretization bias* and its dependence on scale design have not been systematically studied.

Scale Design in Psychometrics Scale design is known to be non-neutral for human respondents: the number of response categories affects reliability and validity (Lozano et al., 2008; Preston and Colman, 2000), and boundary placement biases judgments toward specific anchors (Tversky and Kahneman, 1974; Harvey, 1997). Whether such effects transfer to LLMs is unclear, since their outputs are shaped by token-level distributions, potentially reflecting an additional mechanism absent in humans: the statistical prevalence of numerical tokens in pretraining corpora (Zhao et al., 2021; Shaki et al., 2023).

Metacognitive Sensitivity and Signal Detection Theory Evaluating metacognition requires distinguishing Type-1 performance (task accuracy) from Type-2 sensitivity (the ability to monitor that accuracy) (Galvin et al., 2003). Signal Detection Theory (SDT) provides a principled framework for this decomposition (Hautus et al., 2021). The $meta\text{-}d'$ metric (Maniscalco and Lau, 2012) quantifies metacognitive sensitivity by measuring how well confidence ratings distinguish correct from incorrect answers, independently of response bias (Fleming and Lau, 2014; Rausch et al., 2023). Standard SDT corrections for extreme proportions ensure robust estimation under skewed response distributions (Hautus, 1995). Though recent NLP work has begun disentangling metacognitive ability from task performance (Wang et al., 2025), no study has used SDT-based metrics to examine how confidence scale design influences LLM metacognition. By applying $meta\text{-}d'$ across controlled confidence scales, we address this gap.

3 Methodology

We frame confidence elicitation as a Type-2 (metacognitive) judgment layered on top of a Type-1 (task-level) decision. Given a task instance x from a dataset $\mathcal{D}$, a model M generates an answer $\hat{y}$ and a verbalized confidence score c. The confidence c is constrained within a prompted scale $\mathcal{S} = [l, u]$, where l and u denote the lower and upper bounds respectively, and the model is instructed to report an integer $c \in \{l, l{+}1, \ldots, u\}$. We let $a \in \{0, 1\}$ represent the correctness of $\hat{y}$ against the ground truth y. To enable fair comparison across scales of different ranges, all confidence scores are linearly normalized to the unit interval:

$$\hat{c} = \frac{c - l}{u - l}$$

All subsequent metrics are computed over $\hat{c} \in [0, 1]$. Our central research question is: *how does the metacognitive informativeness of* $\hat{c}$*, defined as its ability to discriminate between* $a{=}1$ *and* $a{=}0$*, vary with the* ***confidence scale*** *design* $\mathcal{S}$*?*

3.1 Scale Design Dimensions

Granularity ($\mathcal{G}$) We vary the granularity of the confidence scale across five levels: $\mathcal{S} \in \{[0, 5],\ [0, 10],\ [0, 20],\ [0, 50],\ [0, 100]\}$. This dimension tests a fundamental trade-off. Finer-grained scales offer higher theoretical resolution

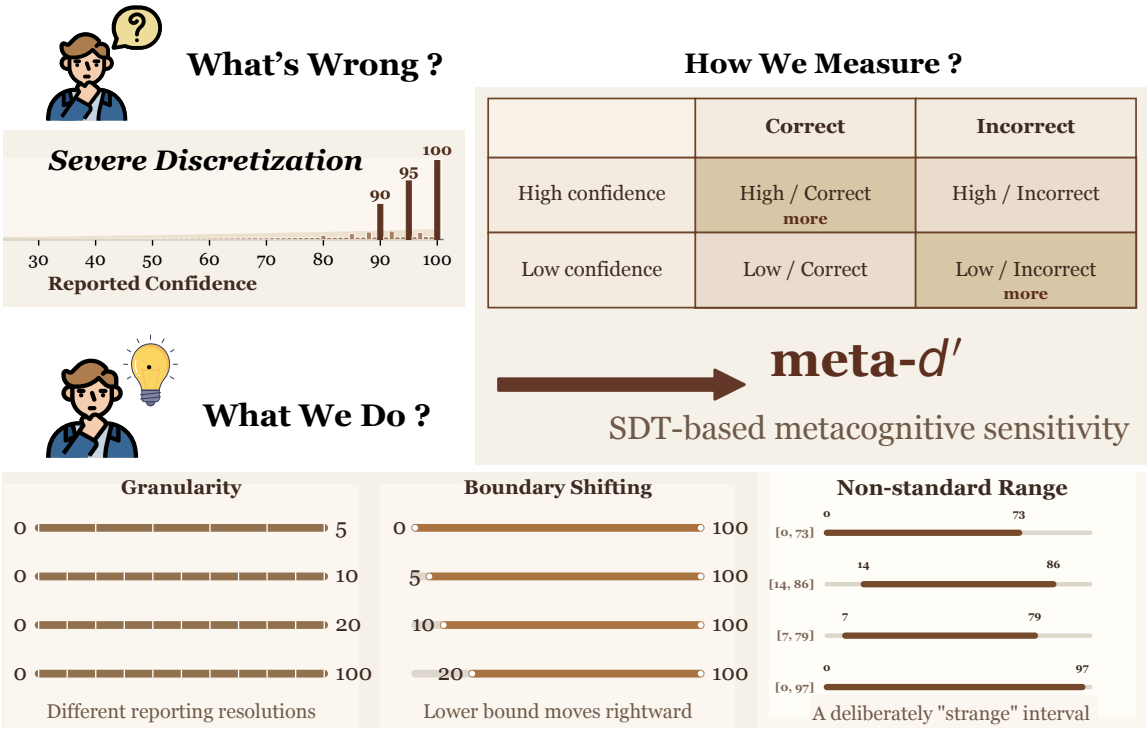

Figure 1: Confidence discretization phenomenon, $meta\text{-}d'$ evaluation, and three-dimensional scale manipulations.

to express uncertainty, but may also amplify token-level biases by providing more opportunities to default to round-number anchors. Conversely, coarser scales force the model to commit to broader confidence bins, potentially reducing noise at the cost of resolution.

Boundary Shifting ($\mathcal{B}$) We fix the upper bound u=100 and progressively raise the lower bound: $\mathcal{S} \in \{[0, 100],\ [20, 100],\ [40, 100],\ [60, 100]\}$. This probes anchoring effects: when the available range is compressed from above, does the model redistribute its confidence to utilize the full new range, or does it remain clustered near the ceiling (Tversky and Kahneman, 1974)? A failure to redistribute would indicate that the model's high-confidence outputs are driven by token preference.

Non-standard Ranges ($\mathcal{N}$) Our range selection follows four principles: *(i)* all bounds avoid conventional multiples of 5 and 10, to eliminate effects from pretraining and encourage reasoning based on scale semantics; *(ii)* both lower and upper bounds vary to avoid relying on a single boundary configuration; *(iii)* include ranges that contain prominent internal round anchors (e.g., 50, 70), while others largely do not, allowing us to test whether models gravitate toward these anchors independently of the scale boundaries; and *(iv)* include a near-standard range to isolate the effect of boundary irregularity from range width.

Concretely, we organize these conditions into two groups. The first holds the range width approximately constant ($\approx$70–73 units) while varying both bounds, isolating the effect of boundary placement: $[0,\ 73]$, $[14,\ 86]$, $[7,\ 79]$. The second group varies the range width to test the robustness of our findings and introduces diagnostic contrasts: $[3,\ 38]$ is narrow with few internal round anchors; $[0,\ 97]$ is a near-standard range (minimal pair with $[0,\ 100]$). If a model instructed to report confidence in $[0,\ 73]$ still clusters at values such as 50 or 70, it indicates reliance on pre-trained numerical heuristics; if reports spread across the admissible range (e.g., 55, 68, 73), it suggests genuine semantic adaptation to the scale constraints. The narrow range $[3,\ 38]$, which contains few salient round-number anchors, provides the strongest diagnostic test of this distinction.

3.2 Evaluation Metrics

Expected Calibration Error (ECE) Following Guo et al. (2017), we partition the normalized confidence space $[0, 1]$ into B equal-width bins and compute:

$$\text{ECE} = \sum_{b=1}^{B} \frac{n_b}{N} \left|\text{acc}(b) - \text{conf}(b)\right|$$

where n_b is the number of samples in bin b, N is the total sample count, and $\text{acc}(b)$ and $\text{conf}(b)$ are the average accuracy and confidence within that bin. We set B=10 as the default and additionally report results with B=15 and adaptive binning in Appendix B, since the choice of B is particularly consequential when confidence distributions are heavily discretized.

AUROC We report the Area Under the Receiver Operating Characteristic curve, which measures the model's ability to assign higher confidence to correct predictions than to incorrect ones (failure prediction). Unlike ECE, AUROC is threshold-free and thus less sensitive to discretization artifacts.

Metacognitive Sensitivity ($meta$-d') To isolate the quality of the metacognitive signal from overall response bias, we adopt $meta$-d' from Signal Detection Theory (Maniscalco and Lau, 2012; Fleming and Lau, 2014). The intuition is as follows: in the Type-1 task, a model's ability to distinguish correct from incorrect answers is captured by $d' = \Phi^{-1}(H_1) - \Phi^{-1}(F_1)$, where $H_1 = P(\hat{c} > t \mid a = 1)$ and $F_1 = P(\hat{c} > t \mid a = 0)$ are the Type-1 hit and false-alarm rates, computed by binarizing $\hat{c}$ at the model's own Type-1 decision criterion. Type-2 (metacognitive) task measures *how well does the model's confidence separate its own correct and incorrect responses?* To formalize this, we binarize $\hat{c}$ at a threshold t to obtain "high confidence" and "low confidence" judgments, then compute:

- *Type-2 Hit Rate*: $H_2 = P(\hat{c} > t \mid a = 1)$, the proportion of correct responses assigned high confidence.
- *Type-2 False Alarm Rate*: $F_2 = P(\hat{c} > t \mid a = 0)$, the proportion of incorrect responses assigned high confidence.

Formally, $meta$-d' is the value satisfying $\Phi^{-1}(H_2) - \Phi^{-1}(F_2) = meta\text{-}d'$, where Φ^{-1} denotes the inverse of the standard normal cumulative distribution function, which converts hit and false-alarm *rates* into z-scores under the equal-variance Gaussian assumption of SDT, where the Type-2 criteria are derived by mapping the Type-1 criterion proportionally onto the $meta$-d' scale (Maniscalco and Lau, 2012). A model with perfect metacognition would have $meta$-$d' = d'$; a model whose confidence carries no information about correctness would have $meta$-$d' = 0$. We further report the metacognitive efficiency ratio $M_{ratio} = meta\text{-}d'/d'$ (Fleming and Lau, 2014), which normalizes for differences in task difficulty across datasets: $M_{ratio} = 1$ indicates that the metacognitive system captures all available Type-1 information, while $M_{ratio} < 1$ indicates metacognitive loss. We use two alternative thresholds ($t = 0.9$ and $t = 0.95$ after normalization) and report results for both. To handle the prevalence of extreme confidence values (e.g., 100%), we apply the log-linear correction of adding 0.5 to all cells of the Type-2 contingency table (Hautus, 1995).

Distribution Diagnostics To quantify discretization artifacts under different scale designs, we introduce two additional diagnostic metrics.

Round measures the proportion of confidence reports that fall on multiples of five (e.g., 0, 5, 10, ..., 100), capturing models' tendency to rely on round-number anchors.

Viol. measures the proportion of responses whose reported confidence lies outside the valid range $[l, u]$ by more than 5% of the scale width.

3.3 Experimental Setup

Models We evaluate six LLMs spanning different training paradigms, model families, and scales.

- **Closed-source:** GPT-5.2 (OpenAI, 2025) and Gemini 3.1 Pro (Google DeepMind, 2025), represent two frontier commercial APIs.
- **Open-weights (Meta):** LLaMA-4-Maverick (17B active / 400B total, 128 experts) and LLaMA-4-Scout (17B active / 109B total, 16 experts) (Meta AI, 2025), enabling within-family comparison across model scales under the same MoE architecture.
- **Open-weights (Alibaba):** Qwen3-235B-A22B-Instruct (22B active / 235B total) and Qwen3-30B-A3B-Instruct (3B active / 30B total) (Yang et al., 2025), broadening coverage to a non-Meta open-weights ecosystem and providing a second within-family scale.

This selection enables three axes of comparison: closed vs. open-source, large vs. small scale, and across model families. We note that all four open-weight models adopt Mixture-of-Experts architectures, and discuss the generalizability implications of this selection in Appendix H. For the scale-granularity analysis, we additionally evaluate Llama-3-8B-Instruct as a dense-architecture robustness baseline. We therefore distinguish the

six primary models used throughout the full experimental design from this seventh model, which appears only in the granularity comparison.

Datasets We select three datasets that span distinct cognitive demands:

- **MMLU** (Hendrycks et al., 2020): 4-choice knowledge-intensive QA covering 57 subjects. We sample a balanced subset of 1,000 instances across difficulty levels.
- **GSM8K** (Cobbe et al., 2021): Multi-step mathematical reasoning. We use the full test set (N=1,319).
- **TruthfulQA** (Lin et al., 2022b): Probes the model's tendency to produce common misconceptions. We use the multiple-choice variant (N=817).

This selection is motivated by the need to test whether scale effects are stable across tasks that vary in difficulty, answer format, and the degree to which overconfidence is a known failure mode.

Prompting We use a standardized zero-shot prompt template: *"Answer the following question. Then, provide your confidence as an integer between [l] and [u], where [l] means no confidence and [u] means absolute certainty."* The explicit anchoring of the endpoints with semantic labels ("no confidence" / "absolute certainty") is held constant across all scale conditions to isolate the effect of numerical range from verbal framing. Full prompt templates are provided in Appendix C.

Inference & Statistical Analysis All inferences are performed with temperature T=0 for deterministic outputs. For each combination of model $\times$ dataset $\times$ scale condition, we collect the full set of available instances ($N \geq 800$), providing sufficient statistical power. To assess the statistical reliability of differences in $meta\text{-}d'$ across scale conditions, we employ a nonparametric bootstrap procedure: we resample the response data 10,000 times with replacement and compute 95% confidence intervals for all metrics. Pairwise comparisons between scale conditions are evaluated via permutation tests (n=10,000), with Bonferroni correction.

4 Results

4.1 Confidence Discretization Under Standard Scales

We first characterize the distribution of verbalized confidence under the standard $[0, 100]$ scale (Table 1; distributions visualized in Figure 2). All six models exhibit severe discretization: models concentrate reports on a sparse set of round-number anchors.

Across models, a single confidence value dominates the distribution. Gemini 3.1 Pro reports exactly 100 on 68.4% of instances, GPT-5.2 and Qwen3-235B most frequently output 95, and the LLaMA-4 family concentrates on 90. The three most common values together account for more than 78% of responses for every model. Notably, this concentration is not simply a byproduct of high task accuracy: conditioned on reporting the Top-1 value, accuracy is substantially lower than the value itself would imply: e.g., Gemini 3.1 Pro is correct on only 79% of the instances where it reports 100, and GPT-5.2 is correct on 85% of instances where it reports 95. This gap indicates that the dominant confidence value functions as a default token-level output (Figure 2(e)). Despite the nominal 101-state scale, models actually use only 15–28 distinct confidence values, leaving most of the scale unused.

Entropy further quantifies this compression. While a uniform distribution over 101 states would yield 6.66 bits, the observed entropy ranges from 0.95 bits (Gemini 3.1 Pro) to 1.88 bits (LLaMA-4-Maverick), indicating highly concentrated confidence distributions.

Baseline metacognitive efficiency (M_{ratio}) varies substantially across models, from 0.62 (Qwen3-30B) to 0.92 (GPT-5.2). A consistent pattern emerges: models with lower entropy (i.e., more concentrated confidence distributions) tend to exhibit lower M_{ratio}. We examine the causal direction of this relationship through controlled scale manipulation experiments in the following sections. AUROC values remain relatively high across models, indicating that even discretized confidence signals still preserve coarse ranking information between correct and incorrect predictions.

4.2 Effect of Scale Granularity

Table 2 reports the effect of scale granularity ($\mathcal{G}$) on metacognitive sensitivity ($meta\text{-}d'$) and efficiency (M_{ratio}). Task accuracy remains stable across granularity conditions, confirming that changes in

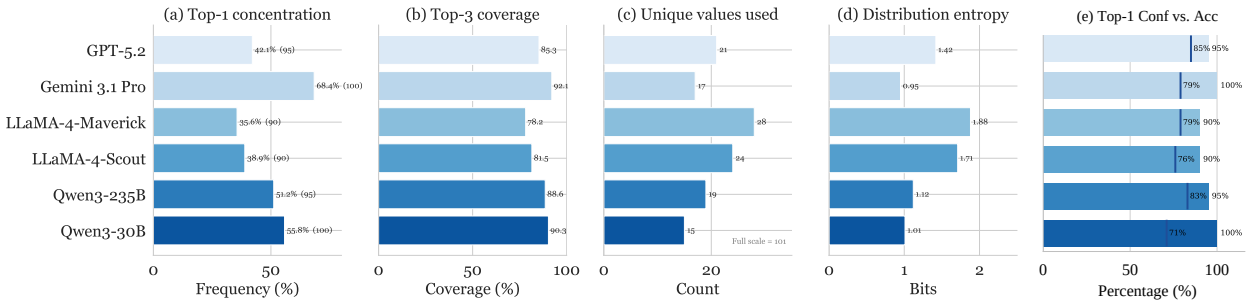

Figure 2: Baseline discretization profile under the standard $[0, 100]$ confidence scale. Across all six models, confidence reports are highly concentrated on a small set of round-number values. The most frequent value alone accounts for 35.6%–68.4% of responses, and the top three values cover 78.2%–92.1%. Models use only 15–28 distinct integers out of 101 possible values, revealing severe discretization of the verbalized confidence signal. Conditional accuracy at the Top-1 confidence value, averaged across datasets in (e) show gaps from 10% to 29%.

	Concentration			Distribution		Performance				
Model	Top-1		Top-3 (%)	Unique	H (bits)	Acc	AUROC	ECE ↓	$meta\text{-}d'$	M_{ratio}
GPT-5.2	95	(42.1%)	85.3	21	1.42	.81	.88	.08	1.84	**0.92**
LLaMA-4-Maverick	90	(35.6%)	78.2	28	1.88	.76	.84	.11	1.65	0.82
Qwen3-235B	95	(51.2%)	88.6	19	1.12	.79	.82	.13	1.51	0.78
LLaMA-4-Scout	90	(38.9%)	81.5	24	1.71	.72	.80	.13	1.52	0.76
Gemini 3.1 Pro	100	(68.4%)	92.1	17	0.95	.78	.79	.15	1.42	0.74
Qwen3-30B	100	(55.8%)	90.3	15	1.01	.68	.74	.17	1.22	0.62

Table 1: Baseline discretization under the standard $[0, 100]$ scale, averaged across MMLU, GSM8K, and TruthfulQA. Models are ordered by M_{ratio} (descending). *Top-1*: most frequent confidence value (frequency in parentheses; bar proportional to frequency). *Top-3*: cumulative coverage of the three most common values. *Unique*: number of distinct confidence integers used out of 101 possible values. *H*: Shannon entropy in bits. *AUROC* for failure prediction. *For reference, a uniform distribution over 101 confidence levels would yield 6.66 bits.*

$meta\text{-}d'$ reflect differences in metacognitive signal quality.

Across the six primary models and the additional dense-architecture baseline, we observe a consistent non-monotonic relationship between scale granularity and metacognitive performance (Figure 3): averaged across models, both $meta\text{-}d'$ and M_{ratio} peak at $[0, 20]$ and decline toward both extremes. At the coarse end, $[0, 5]$ yields the lowest $meta\text{-}d'$ for every model, indicating that five categories are insufficient to capture meaningful confidence variation. At the fine end, the standard $[0, 100]$ scale does not outperform $[0, 20]$ for any model; every model achieves significantly higher M_{ratio} at $[0, 20]$ than at $[0, 100]$ ($p < 0.05$).

This pattern holds across both closed-source models (e.g., GPT-5.2 $M_{ratio} = 0.95$ at $[0, 20]$ vs. 0.92 at $[0, 100]$) and open-weight models (e.g., LLaMA-4-Maverick: 0.89 vs. 0.82), suggesting that the $[0, 20]$ advantage is robust to architectural and training differences. Llama-3-8B-Instruct, a dense architecture model, exhibits the same non-monotonic pattern ($M_{ratio} = 0.79$ at $[0, 20]$ vs. 0.71 at $[0, 100]$, $p < 0.05$), confirming that the finding

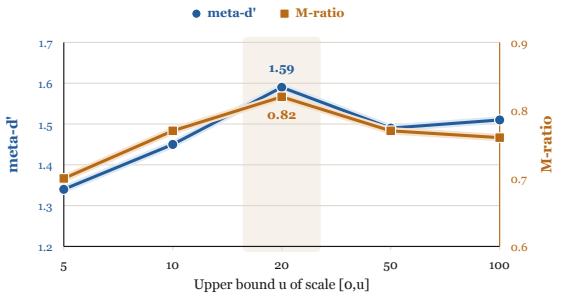

Figure 3: Averaged $meta\text{-}d'$ and M_{ratio} across 7 models along scale granularity. Both metrics peak at $[0, 20]$ and decline toward both coarser and finer extremes.

is not specific to Mixture-of-Experts architectures. Per-dataset breakdowns are in Appendix D.3.

4.3 Effect of Boundary Shifting

Table 3 reports the results of the boundary shifting experiment ($\mathcal{B}$), where the upper bound is fixed at 100 while the lower bound is progressively increased. Two patterns deserve attention.

First, moderate boundary shifts have minimal impact. Moving from $[0, 100]$ to $[20, 100]$ does not produce significant changes in $meta\text{-}d'$ or M_{ratio}

Model	$meta\text{-}d' \uparrow$					$M_{ratio} \uparrow$				
	[0,5]	[0,10]	[0,20]	[0,50]	[0,100]	[0,5]	[0,10]	[0,20]	[0,50]	[0,100]
GPT-5.2	1.68	1.75	**1.92***	1.80	1.84	0.84	0.90	**0.95***	0.91	0.92
Gemini 3.1 Pro	1.30	1.40	**1.52***	1.44	1.42	0.68	0.73	**0.79***	0.75	0.74
LLaMA-4-Maverick	1.41	1.55	**1.73***	1.60	1.65	0.78	0.85	**0.89***	0.81	0.82
LLaMA-4-Scout	1.28	1.40	**1.55***	1.42	1.52	0.71	0.78	**0.83***	0.77	0.76
Qwen3-235B	1.39	1.53	**1.62***	1.55	1.51	0.72	0.79	**0.84***	0.80	0.78
Qwen3-30B	1.08	1.20	**1.34***	1.24	1.22	0.55	0.61	**0.68***	0.63	0.62
Llama-3-8B-Instruct	1.22	1.35	**1.48***	1.38	1.40	0.63	0.70	**0.79***	0.72	0.71

Table 2: Effect of scale granularity ($\mathcal{G}$) on metacognitive sensitivity ($meta\text{-}d'$) and efficiency (M_{ratio}), averaged across datasets. Llama-3-8B-Instruct is included as a dense architecture baseline. Shaded cells indicate the best value per model; * denotes significant improvement over the $[0, 100]$ baseline ($p < 0.05$, permutation test with Bonferroni correction). Confidence values are normalized to $[0, 1]$ prior to metric computation.

for any models, and utilization (*Util.*) increases slightly for several models, suggesting that models can partially adapt to a mildly compressed range. Violation rates under all standard and boundary-shifted conditions are at or near zero (see Table 21).

Second, aggressive boundary shifts lead to consistent degradation. In the $[40, 100]$ condition, M_{ratio} decreases for most models, and in the $[60, 100]$ condition the drop is universal and substantial. The Utilization metric reveals the mechanism: as the lower bound increases, models fail to redistribute their confidence across the new range and instead cluster more tightly near the upper bound (under the $[60, 100]$ condition Qwen3-30B uses only 3.1% of the available range).

4.4 Semantic Robustness of Non-standard

Tables 4 and 5 present results for the non-standard scale experiment ($\mathcal{N}$), designed to disentangle genuine scale comprehension from round-number.

Group 1: Fixed width, varying bounds Under the three fixed-width conditions ($[0, 73]$, $[14, 86]$, $[7, 79]$; Table 4), Round (*round-number preference*) decreases relative to the baseline $[0, 100]$ scale but remains high. This indicates that irregular boundaries weaken but do not eliminate round-number clustering. The three Group-1 conditions produce nearly identical results (M_{ratio} within ± 0.01 across models), suggesting that the observed effects are driven by scale irregularity.

Group 2: Diagnostic contrasts Table 5 reports results for two contrastive conditions. The narrow $[3, 38]$ scale produces the strongest disruption: violation rates increase substantially (e.g., up to 21.2% for Qwen3-30B), indicating that many responses fall outside the instructed range. Metacognitive efficiency also declines (M_{ratio} 0.40–0.72), while

Model	Metric	Scale Condition [0,100]	[20,100]	[40,100]	[60,100]
GPT-5.2	$meta\text{-}d'$	1.84	1.80	1.62	1.38
	M_{ratio}	0.92	0.90	0.81†	0.69†
	Util.	12.4	14.1	10.8	7.2
Gemini 3.1 Pro	$meta\text{-}d'$	1.42	1.45	1.28	1.12
	M_{ratio}	0.74	0.76	0.67	0.58†
	Util.	6.9	8.6	6.1	4.8
LLaMA-4 Maverick	$meta\text{-}d'$	1.65	1.61	1.48	1.25
	M_{ratio}	0.82	0.80	0.74†	0.63†
	Util.	10.5	11.8	8.9	5.6
LLaMA-4 Scout	$meta\text{-}d'$	1.52	1.49	1.35	1.18
	M_{ratio}	0.76	0.74	0.68†	0.59†
	Util.	9.2	10.1	7.5	4.9
Qwen3 235B	$meta\text{-}d'$	1.51	1.48	1.32	1.05
	M_{ratio}	0.78	0.76	0.68†	0.55†
	Util.	9.1	7.4	6.2	3.9
Qwen3 30B	$meta\text{-}d'$	1.22	1.18	1.02	0.85
	M_{ratio}	0.62	0.60	0.52†	0.43†
	Util.	7.8	6.5	4.8	3.1

Table 3: Effect of boundary shifting ($\mathcal{B}$) on metacognitive performance, averaged across datasets. *Util.*: percentage of the available range covered by the central 90% of reports (5th–95th percentile), reducing sensitivity to isolated outliers. Cell shading indicates direction of change relative to $[0, 100]$: improve, moderate†, severe† (†: $p < 0.05$, permutation test).

Round-Pref remains above 58% for all models, suggesting persistent preference for multiples of five even when few salient anchors exist.

The minimal pair $[0, 97]$ vs. $[0, 100]$ provides a targeted diagnostic. Despite a difference of only three units in the upper bound, M_{ratio} decreases for every model and Round drops slightly. This consistent shift suggests that the value "100" functions as a particularly strong confidence anchor.

These findings suggest that the *discretization*

Model	Metric	Scale Condition [0,100]	[0,73]	[14,86]	[7,79]
GPT-5.2	Round (%)	94.2	76.8	72.1	74.5
	Viol. (%)	0.0	1.8	2.4	2.1
	$meta\text{-}d'$	1.84	1.71	1.68	1.70
	M_{ratio}	0.92	0.85	0.84	0.85
Gemini 3.1	Round (%)	96.5	82.4	79.6	80.8
	Viol. (%)	0.0	4.1	5.2	4.5
	$meta\text{-}d'$	1.42	1.28	1.24	1.26
	M_{ratio}	0.74	0.67	0.65	0.66
LLaMA-4-M	Round (%)	88.7	71.2	68.5	69.8
	Viol. (%)	0.0	2.6	3.1	2.8
	$meta\text{-}d'$	1.65	1.52	1.49	1.50
	M_{ratio}	0.82	0.76	0.74	0.75
LLaMA-4-S	Round (%)	90.1	74.8	71.9	73.2
	Viol. (%)	0.0	3.5	4.2	3.8
	$meta\text{-}d'$	1.52	1.38	1.35	1.36
	M_{ratio}	0.76	0.69	0.68	0.68
Qwen3-235B	Round (%)	92.1	78.4	75.2	76.8
	Viol. (%)	0.0	3.2	4.0	3.5
	$meta\text{-}d'$	1.51	1.38	1.35	1.36
	M_{ratio}	0.78	0.71	0.70	0.70
Qwen3-30B	Round (%)	95.3	82.6	80.1	81.2
	Viol. (%)	0.0	5.8	6.5	6.1
	$meta\text{-}d'$	1.22	1.05	1.02	1.03
	M_{ratio}	0.62	0.53	0.52	0.52

Table 4: Diagnostics under fixed-width non-standard scales (Group 1, width ≈72–73), averaged across datasets. *Round*: reports on multiples of 5 within $[l, u]$. *Viol.*: responses outside valid range by >5% of width. Cell shading: reduced Round , meta. decline , elevated Viol. , all relative to $[0, 100]$.

Model	Metric	Scale Condition [0,100]	[3,38]	[0,97]
GPT-5.2	Round (%)	94.2	58.3	89.1
	Viol. (%)	0.0	8.6	0.5
	$meta\text{-}d'$	1.84	1.45	1.79
	M_{ratio}	0.92	0.72	0.90
Gemini 3.1	Round (%)	96.5	65.2	93.8
	Viol. (%)	0.0	18.3	0.8
	$meta\text{-}d'$	1.42	0.95	1.38
	M_{ratio}	0.74	0.50	0.72
LLaMA-4-M	Round (%)	88.7	59.4	85.3
	Viol. (%)	0.0	12.9	0.4
	$meta\text{-}d'$	1.65	1.28	1.61
	M_{ratio}	0.82	0.64	0.80
LLaMA-4-S	Round (%)	90.1	62.1	87.5
	Viol. (%)	0.0	15.4	0.6
	$meta\text{-}d'$	1.52	1.14	1.48
	M_{ratio}	0.76	0.57	0.74
Qwen3-235B	Round (%)	92.1	62.5	89.5
	Viol. (%)	0.0	14.8	0.6
	$meta\text{-}d'$	1.51	1.12	1.47
	M_{ratio}	0.78	0.58	0.76
Qwen3-30B	Round (%)	95.3	68.4	92.8
	Viol. (%)	0.0	21.2	1.2
	$meta\text{-}d'$	1.22	0.78	1.18
	M_{ratio}	0.62	0.40	0.60

Table 5: Diagnostics under contrastive non-standard scales (Group 2), averaged across datasets. $[3, 38]$: narrow range, few internal round anchors (stress test). $[0, 97]$: minimal pair with $[0, 100]$ (3-unit shift). Cell shading: reduced Round , mod. decline , severe decline , high Viol.

patterns observed under standard scales are not merely artifacts of specific numeric boundaries, but reflect persistent anchor-seeking behavior in verbalized confidence generation.

4.5 Temperature Robustness

To assess whether the $[0, 20]$ advantage is an artifact of greedy decoding, we evaluate GPT-5.2 and LLaMA-4-Maverick on MMLU under stochastic sampling at T=0.3 and T=1. Round-number preference decreases with temperature but persists even at T=1 (GPT-5.2: 88.1% under $[0, 100]$), consistent with anchor-seeking behavior driven by pretraining token frequencies. Critically, ΔM_{ratio}, the improvement of $[0, 20]$ over $[0, 100]$, remains identical across all three temperature conditions for both models (+0.03 for GPT-5.2, +0.07 for LLaMA-4-Maverick), confirming that the $[0, 20]$ advantage is not an artifact of deterministic decoding. Full results are provided in Appendix E.

4.6 Prompt Variation

To assess whether the $[0, 20]$ advantage depends on our specific prompt wording, we evaluate GPT-5.2 on MMLU under two additional conditions: role prompting (calibrated forecaster persona) and negation framing (reversed endpoint labels). Role prompting produces negligible changes in M_{ratio} relative to baseline, while negation framing introduces a small uniform drop under both scale conditions. In all cases, ΔM_{ratio} between $[0, 20]$ and $[0, 100]$ remains +0.03, confirming that the finding is robust to prompt variation. Full results are provided in Appendix F.

5 Conclusion

To our knowledge, we present the first systematic study of confidence scale design as an experimental variable in LLM evaluation. Across six primary models and three datasets, with an additional dense-architecture model used for the granularity robust-

ness analysis, we show that verbalized confidence is severely discretized under the standard $[0, 100]$ scale. Within the evaluated settings, a $[0, 20]$ scale consistently yields higher metacognitive efficiency, aggressive boundary compression degrades performance, and round-number preference persists even under irregular ranges. Based on these findings, we recommend three practices: *(i)* use a $[0, 20]$ scale as an alternative to the standard $[0, 100]$ format; *(ii)* report $meta\text{-}d'$ alongside ECE, since ECE becomes unreliable under highly discretized distributions; and *(iii)* inspect empirical confidence distributions before interpreting calibration metrics. Confidence scale design is often treated as a neutral reporting choice, our results show it is not.

Limitations and Future Work

All experiments use greedy decoding ($T{=}0$); temperature robustness is validated in Appendix E. We use zero-shot prompting without Chain-of-Thought or self-reflection steps; prompt variation results are reported in Appendix F. Our closed-source models have undisclosed architectures, while open-weight models use Mixture-of-Experts architectures; dense model results are included in Table 2. We evaluate on three datasets covering knowledge QA, mathematical reasoning, and misconception detection; generalization to open-ended generation tasks is not established. Finally, $meta\text{-}d'$ requires binarizing confidence at a threshold; a continuous Type-2 ROC analysis could provide a more comprehensive characterization and is left to future work.

Ethics Statement

This work uses only publicly available benchmark datasets. All model outputs are generated through standard API access or publicly released model weights. We do not anticipate direct negative societal impacts from this research. However, we note that our findings suggest that confidence scores are shaped by scale design rather than solely reflecting genuine model uncertainty. This has implications for deployed systems that present verbalized confidence to end users, since overconfidence driven by scale artifacts could lead to misplaced trust. We encourage practitioners to adopt our recommendations when designing confidence elicitation interfaces.

References

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Stephen M Fleming and Hakwan C Lau. 2014. How to measure metacognition. *Frontiers in human neuroscience*, 8:82285.

Susan J Galvin, John V Podd, Vit Drga, and John Whitmore. 2003. Type 2 tasks in the theory of signal detectability: Discrimination between correct and incorrect decisions. *Psychonomic bulletin & review*, 10(4):843–876.

Google DeepMind. 2025. Gemini 3: Introducing the latest gemini ai model from google.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Nigel Harvey. 1997. Confidence in judgment. *Trends in cognitive sciences*, 1(2):78–82.

Michael J Hautus. 1995. Corrections for extreme proportions and their biasing effects on estimated values of d'. *Behavior research methods, instruments, & computers*, 27(1):46–51.

Michael J Hautus, Neil A Macmillan, and C Douglas Creelman. 2021. *Detection theory: A user's guide*. Routledge.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Ananya Kumar, Percy S Liang, and Tengyu Ma. 2019. Verified uncertainty calibration. *Advances in neural information processing systems*, 32.

Loka Li, Zhenhao Chen, Guangyi Chen, Yixuan Zhang, Yusheng Su, Eric Xing, and Kun Zhang. 2024. Confidence matters: Revisiting intrinsic self-correction capabilities of large language models. *arXiv preprint arXiv:2402.12563*.

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

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022b. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 3214–3252.

Luis M Lozano, Eduardo García-Cueto, and José Muñiz. 2008. Effect of the number of response categories on the reliability and validity of rating scales. *Methodology*, 4(2):73–79.

Brian Maniscalco and Hakwan Lau. 2012. A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and cognition*, 21(1):422–430.

Meta AI. 2025. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation.

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

OpenAI. 2025. Introducing gpt-5.

Max Pellert, Clemens M Lechner, Claudia Wagner, Beatrice Rammstedt, and Markus Strohmaier. 2024. Ai psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*, 19(5):808–826.

Carolyn C Preston and Andrew M Colman. 2000. Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta psychologica*, 104(1):1–15.

Manuel Rausch, Sebastian Hellmann, and Michael Zehetleitner. 2023. Measures of metacognitive efficiency across cognitive models of decision confidence. *Psychological Methods*.

Jonathan Shaki, Sarit Kraus, and Michael Wooldridge. 2023. Cognitive effects in large language models. *arXiv preprint arXiv:2308.14337*.

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

Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131.

Guoqing Wang, Wen Wu, Guangze Ye, Zhenxiao Cheng, Xi Chen, and Hong Zheng. 2025. Decoupling metacognition from cognition: a framework for quantifying metacognitive ability in llms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25353–25361.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Yuqing Wang and Yun Zhao. 2024. Metacognitive prompting improves understanding in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1914–1926.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

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. arXiv*, 2306.

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

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pages 12697–12706. Pmlr.

Kaitlyn Zhou, Jena Hwang, Xiang Ren, and Maarten Sap. 2024. Relying on the unreliable: The impact of language models' reluctance to express uncertainty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3623–3643.

Appendix Overview

A Related Work

A.1 Verbalized Confidence and Calibration

As LLMs increasingly became available through black-box commercial APIs, research on uncertainty estimation shifted from logit-based methods (Kadavath et al., 2022) toward verbalized confidence, where models are prompted to express certainty as a numerical score (Lin et al., 2022a; Xiong et al., 2024). Substantial effort has focused on improving confidence elicitation through prompting and inference strategies, including Chain-of-Thought prompting (Wei et al., 2022), metacognitive prompting (Wang and Zhao, 2024), self-consistency aggregation (Wang et al., 2022), and techniques tailored to RLHF-aligned models (Tian et al., 2023). Evaluation has likewise centered on calibration metrics, predominantly Expected Calibration Error (ECE: Guo et al., 2017). Yet this literature largely treats the confidence scale as a fixed backdrop, assuming that a 0–100 scale faithfully conveys whatever metacognitive signal the model possesses.

However, we find that LLMs do not use the 0–100 scale as a continuous spectrum but cluster outputs around a small set of round-number anchors. Then, when models concentrate the vast majority of their reports on two or three values, ECE bins become sparse or empty across most of the range, making calibration estimates unreliable. This *discretization bias* and its dependence on scale design have not been systematically studied.

A.2 Scale Design in Psychometrics

In human subject research, scale design is known to be non-neutral: the number of response categories modulates both reliability and validity (Lozano et al., 2008; Preston and Colman, 2000), and scale boundaries bias judgments toward specific anchors (Tversky and Kahneman, 1974; Harvey, 1997). While such sensitivities are well-documented for human respondents, it remains unclear whether they apply to LLMs, whose outputs are driven by token-level distributions rather than human cognition. LLMs may exhibit analogous effects, or their biases may reflect an additional mechanism not present in human respondents: the statistical prevalence of certain numerical tokens in pre-training corpora (Zhao et al., 2021; Shaki et al., 2023).

A.3 Metacognitive Sensitivity and Signal Detection Theory

Evaluating metacognition requires distinguishing Type-1 performance (task accuracy) from Type-2 sensitivity (the ability to monitor that accuracy) (Galvin et al., 2003). Signal Detection Theory (SDT) provides a principled framework for this decomposition (Hautus et al., 2021). The $meta\text{-}d'$ metric (Maniscalco and Lau, 2012) quantifies metacognitive sensitivity by measuring how well confidence ratings distinguish correct from incorrect answers, independently of response bias (Fleming and Lau, 2014; Rausch et al., 2023). Standard SDT corrections for extreme proportions ensure robust estimation under skewed response distributions (Hautus, 1995). Though recent NLP work has begun disentangling metacognitive ability from task performance (Wang et al., 2025), no study has used SDT-based metrics to examine how confidence scale design influences LLM metacognition. By applying $meta\text{-}d'$ across controlled confidence scales, we address this gap.

B Sensitivity of ECE to Bin Count and Binning Strategy

A known limitation of Expected Calibration Error is its dependence on the number of bins B and the binning strategy (Nixon et al., 2019; Kumar et al., 2019). This sensitivity is particularly acute in our setting, where confidence distributions are heavily discretized—the majority of probability mass is concentrated on two or three values (e.g., 95 and 100 under the $[0, 100]$ scale). Under such distributions, most equal-width bins are empty or near-empty, and small changes in B can shift samples across bin boundaries, producing unstable ECE estimates.

B.1 Effect of B on ECE Estimates

Figure 4 shows how ECE varies as a function of $B \in \{5, 10, 15, 20, 30, 50\}$ for a representative model–dataset pair (GPT-5.2 on MMLU) under the standard $[0, 100]$ scale. ECE fluctuates noticeably across bin counts, with relative differences of up to approximately 13% between B=10 and B=20. This instability arises because the model's confidence reports are highly discretized and concentrated on a small set of values (e.g., 90, 95, and 100). After normalization to the $[0, 1]$ interval, most probability mass falls within a narrow region of the scale, causing bin membership to be highly sensitive to boundary placement. As a result, small changes in B can reassign a substantial number of observations across bins, producing noticeable variation in the estimated calibration error.

B.2 Equal-Width vs. Equal-Mass Binning

The standard ECE formulation uses *equal-width* bins, which partition $[0, 1]$ into B intervals of size $1/B$. Under discretized confidence, this produces a highly uneven distribution of samples across bins. An alternative is *equal-mass* (or *quantile*) binning, which assigns approximately N/B samples to each bin, ensuring that no bin is empty (Nixon et al., 2019).

Table 6 reports ECE under both strategies with B=10 for representative model–scale combinations on MMLU. We highlight two patterns. First, under the standard $[0, 100]$ scale, the discrepancy between equal-width and equal-mass ECE is substantial, reaching up to 0.037 absolute points (e.g., for LLaMA-4-Maverick). This confirms that binning artifacts can dominate the calibration estimate when confidence values are concentrated on a small set of anchors. Second, under the coarser $[0, 20]$ scale, the discrepancy shrinks considerably (typically below 0.01), indicating that the distribution of confidence values becomes more evenly spread across bins. Importantly, these differences arise despite identical model predictions and confidence reports, indicating that the observed variation is purely an artifact of the binning procedure. This result supports the interpretation that scale design directly shapes the empirical confidence distribution and, consequently, the stability of calibration metrics.

Table 6: Comparison of ECE under **Equal-Width** vs. **Equal-Mass** binning ($B = 10$) on MMLU. Large $|\Delta|$ indicates metrics driven by discretization artifacts.

Model	Scale	Equal-Width	Equal-Mass	$\lvert\Delta\rvert$
GPT-5.2	[0, 100]	0.082	0.051	0.031
	[0, 20]	0.054	0.048	0.006
LLaMA-4-Mav.	[0, 100]	0.115	0.078	0.037
	[0, 20]	0.082	0.075	0.007

B.3 Complete ECE Results Across Bin Counts

For completeness, Table 7 reports ECE with $B \in \{10, 15, 20\}$ for all experimental conditions (model $\times$ dataset $\times$ scale). Our main findings are robust to the choice of B: although absolute ECE values vary as the number of bins changes, the relative ordering

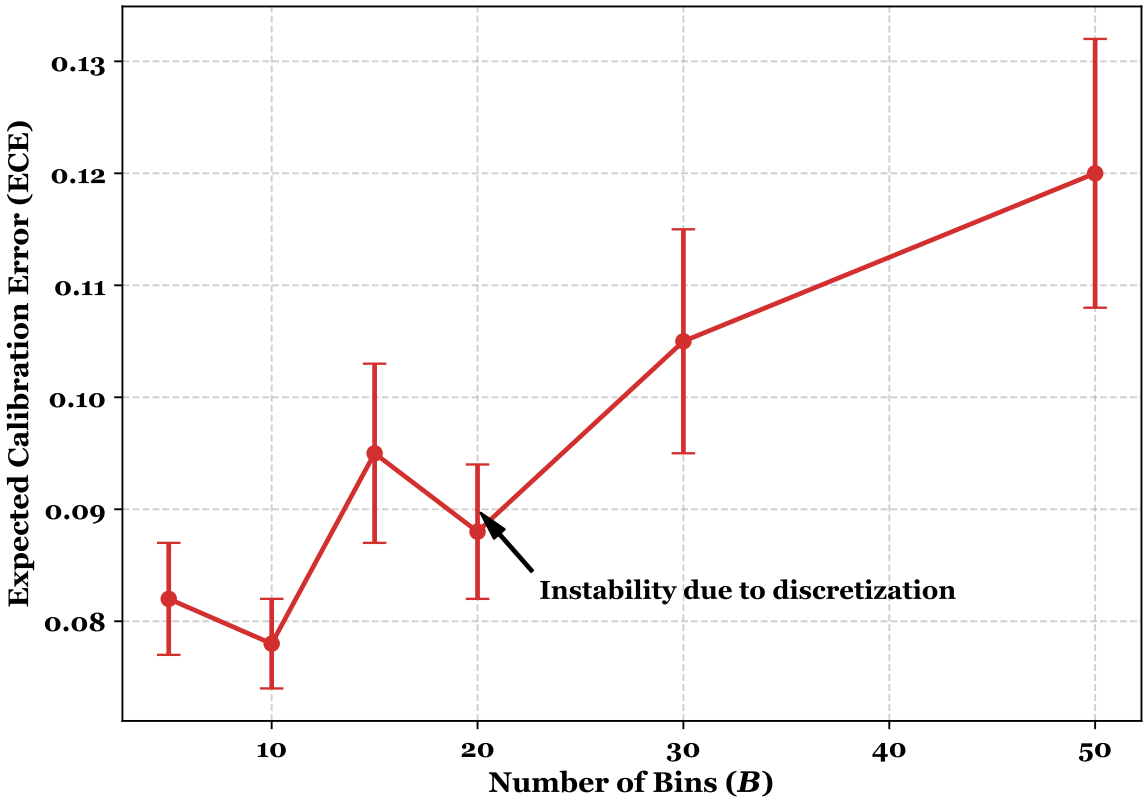

Figure 4: Sensitivity of Expected Calibration Error (ECE) to the number of bins B under discretized confidence distributions. When confidence values concentrate on a small set of anchors, small changes in B can shift observations across bin boundaries, producing unstable ECE estimates.

of scale conditions remains stable across all tested bin counts. In particular, the $[0, 20]$ scale consistently yields the lowest ECE values, followed by $[0, 50]$, while the standard $[0, 100]$ scale produces the highest ECE across all settings.

This stability in ranking indicates that our conclusions about scale design are not driven by a specific choice of bin count. Instead, the differences across scales reflect systematic changes in the empirical confidence distribution. At the same time, the noticeable variation in absolute ECE values across B further illustrates the sensitivity of bin-based calibration metrics under discretized confidence distributions, reinforcing the motivation for using $meta\text{-}d'$ as a complementary, distribution-agnostic measure of metacognitive sensitivity.

B.4 Confidence Distribution Histograms

Figures 5 and 6 visualize the raw (pre-normalization) confidence distributions for each scale condition, aggregated across datasets. These histograms directly illustrate the discretization phenomenon discussed in Section 4 and motivate our use of $meta\text{-}d'$ as a distribution-agnostic alternative to ECE.

These visualizations serve three purposes: (1) they provide direct evidence of discretization as a function of scale design; (2) they explain *why* ECE estimates are unstable—when 80%+ of samples fall in a single bin, the metric is dominated by that bin's accuracy-confidence gap; and (3) they visually preview the key result that coarser and non-standard scales can reshape the confidence distribution in ways that are informative for analysis in the main text.

C Prompt Templates

All prompts follow a standardized two-part structure: a *task instruction* that presents the question and elicits an answer, followed by a *confidence instruction* that specifies the scale and requests a numerical self-assessment. The task instruction is held constant across all scale conditions; only the confidence instruction varies. This design isolates the effect of scale design from any confounds introduced by prompt wording.

Table 7: ECE across experimental conditions with $B \in \{10, 15, 20\}$. Ranking of scales is stable ($[0, 20] < [0, 50] < [0, 100]$) across different B values.

	GPT-5.2 (MMLU)			LLaMA-4-Mav (MMLU)		
Scale	$B = 10$	$B = 15$	$B = 20$	$B = 10$	$B = 15$	$B = 20$
$[0, 20]$	**0.051**	**0.054**	**0.058**	**0.082**	**0.085**	**0.089**
$[0, 50]$	0.078	0.082	0.085	0.105	0.112	0.118
$[0, 100]$	0.082	0.095	0.102	0.115	0.132	0.145

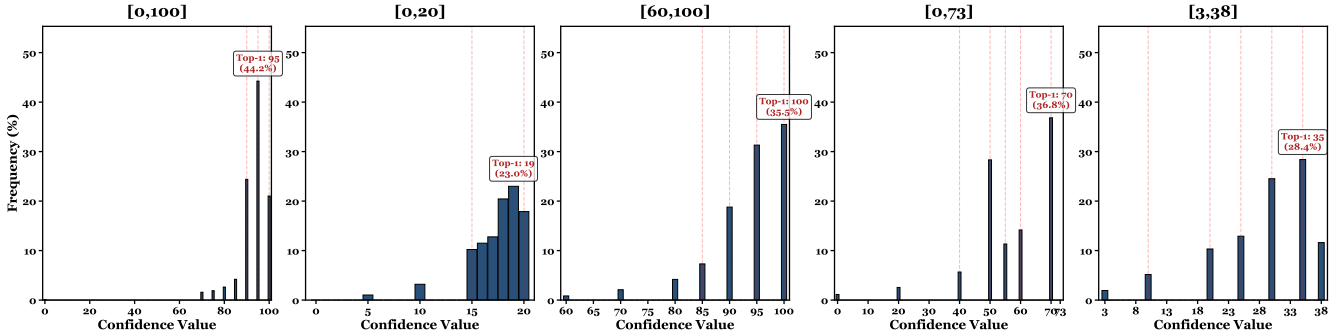

Figure 5: Distribution of raw confidence values across representative scale conditions for GPT-5.2, aggregated across datasets. Under the standard $[0, 100]$ scale, confidence reports concentrate on a small set of round-number anchors, whereas alternative scale designs produce qualitatively different distributional patterns. Dashed vertical lines indicate the most frequent round-number anchors within each range.

C.1 General Template

```
[TASK INSTRUCTION]

Answer the following question.
{question}
Choices: {choices}

[CONFIDENCE INSTRUCTION]
After providing your answer, rate your
confidence as an integer between {l} and
{u}, where {l} means no confidence at all
and {u} means absolute certainty.

[OUTPUT FORMAT]
Respond in the following format only:
Answer: <your answer>
Confidence: <integer between {l} and
{u}>
```

The explicit anchoring of the endpoints with semantic labels ("no confidence at all" / "absolute certainty") is held constant across all conditions. The strict output format minimizes post-processing ambiguity and discourages the model from hedging or elaborating on its confidence.

For GSM8K (free-form numerical answers), the task instruction is adapted as follows:

```
Solve the following math problem step by
step.
{question}

Respond in the following format only:
Solution: <your step-by-step solution>
Final Answer: <numerical answer>
Confidence: <integer between {l} and
{u}>
```

C.2 Scale-Specific Confidence Instructions

Below we list the exact confidence instruction for each scale condition. All other parts of the prompt remain identical.

C.2.1 Granularity Conditions ($\mathcal{G}$)

[0, 5]: `Rate your confidence as an integer between 0 and 5, where 0 means no confidence at all and 5 means absolute certainty.`

[0, 10]: `Rate your confidence as an integer between 0 and 10, where 0 means no confidence at all and 10 means absolute certainty.`

[0, 20]: `Rate your confidence as an integer between 0 and 20, where 0 means no confidence at all and 20 means absolute certainty.`

[0, 50]: `Rate your confidence as an integer between 0 and 50, where 0 means no confidence at all and 50 means absolute certainty.`

[0, 100]: `Rate your confidence as an integer between 0 and 100, where 0 means no confidence at all and 100 means absolute certainty.`

C.2.2 Boundary Shifting Conditions ($\mathcal{B}$)

[0, 100]: (Same as granularity baseline above.)

[20, 100]: `Rate your confidence as an integer between 20 and 100, where 20 means no confidence at all and 100 means absolute certainty.`

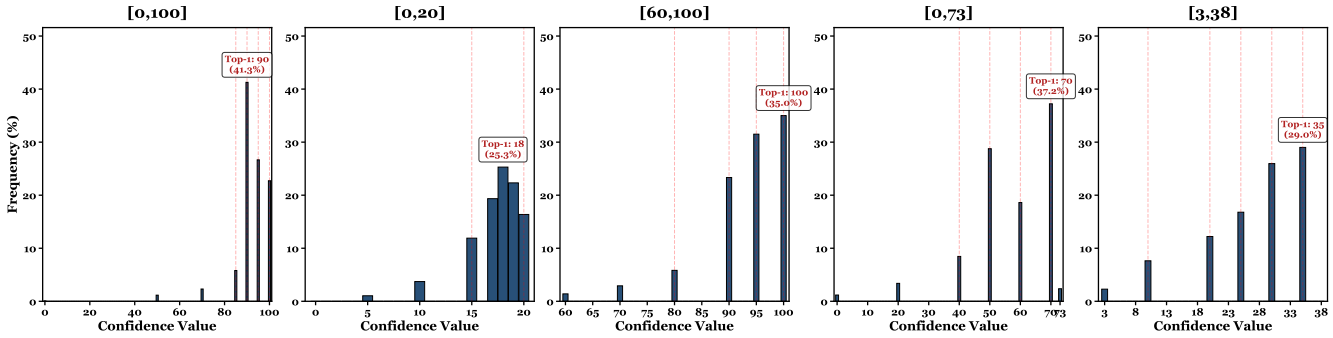

Figure 6: Distribution of raw confidence values across representative scale conditions for LLaMA-4-Maverick, aggregated across datasets. Similar to GPT-5.2, the standard $[0, 100]$ scale produces strong clustering on round-number anchors, while coarser, shifted, and non-standard scales alter the distributional pattern. Dashed vertical lines indicate the most frequent round-number anchors within each range.

> **[40, 100]:** `Rate your confidence as an integer between 40 and 100, where 40 means no confidence at all and 100 means absolute certainty.`
>
> **[60, 100]:** `Rate your confidence as an integer between 60 and 100, where 60 means no confidence at all and 100 means absolute certainty.`

C.2.3 Non-standard Range Conditions ($\mathcal{N}$)

Group 1: Fixed width ($\approx$72–73 units), varying bounds.

> **[0, 73]:** `Rate your confidence as an integer between 0 and 73, where 0 means no confidence at all and 73 means absolute certainty.`
>
> **[14, 86]:** `Rate your confidence as an integer between 14 and 86, where 14 means no confidence at all and 86 means absolute certainty.`
>
> **[7, 79]:** `Rate your confidence as an integer between 7 and 79, where 7 means no confidence at all and 79 means absolute certainty.`

Group 2: Varying width, diagnostic contrasts.

> **[3, 38]:** `Rate your confidence as an integer between 3 and 38, where 3 means no confidence at all and 38 means absolute certainty.`
>
> **[0, 97]:** `Rate your confidence as an integer between 0 and 97, where 0 means no confidence at all and 97 means absolute certainty.`

C.3 Design Rationale and Controls

Several design choices warrant explicit justification:

Semantic anchoring. Every confidence instruction anchors the lower bound to "no confidence at all" and the upper bound to "absolute certainty," regardless of the numerical values. This ensures that any observed differences across scale conditions are attributable to the numerical range itself, not to differences in verbal framing. Without this control, a model might interpret "20" in a $[20, 100]$ scale as moderate confidence rather than as the minimum, confounding boundary shifting effects with semantic reinterpretation.

Integer constraint. We explicitly instruct the model to provide an integer. This prevents the model from reporting values like "95.5" or "around 90," which would introduce parsing ambiguity. In pilot experiments, omitting this constraint led to approximately 8% of responses containing non-integer or range-valued confidence reports (e.g., "85–90").

Output format enforcement. The strict `Answer: / Confidence:` format serves two purposes: (1) it enables reliable automated extraction of confidence values, and (2) it discourages the model from embedding caveats or qualifications that might implicitly modulate the confidence signal (e.g., "I'm fairly confident, so 85"). Responses that do not conform to the specified format are excluded from analysis. Violation rates for each scale condition are reported in Table 21.

No Chain-of-Thought for confidence. We deliberately do not prompt the model to "explain your confidence" or "think step by step about your certainty." While such prompts might improve calibration (Xiong et al., 2024; Tian et al., 2023), they would introduce an additional variable that interacts with scale design. Our goal is to measure the raw effect of scale manipulation on the default confidence signal; the interaction between prompting

strategies and scale design is left for future work.

D Full Experimental Results

This appendix provides the complete experimental results underlying the analyses in Section 4. Main-text tables report averages across datasets and, for selected results, only representative subsets due to space constraints. Here we provide per-dataset baselines, per-dataset conditional accuracy gaps, per-dataset granularity breakdowns, full-model comparisons, and statistical confidence intervals.

D.1 Per-Dataset Baseline Statistics

Tables 8, 9, and 10 report the baseline discretization statistics from Section 4 separately for each dataset. This decomposition verifies that the severe confidence discretization observed in the averaged results is consistent across knowledge-intensive QA, mathematical reasoning, and misconception-probing tasks.

D.2 Per-Dataset Conditional Accuracy Gap

Tables 11, 12, and 13 report the Top-1 conditional accuracy gap (cf. Figure 2(e) in the main text) separately for each dataset, verifying that the overconfidence pattern at the dominant confidence anchor is not driven by a single task type. Averaging each model's per-dataset gap recovers the values reported in the main text (e.g., GPT-5.2: $(6 + 10 + 14)/3 = 10$ pts).

D.3 Granularity: ECE and Per-Dataset Breakdown

Table 14 reports ECE across all granularity conditions, complementing the $meta\text{-}d'$ and M_{ratio} results in the main text. Table 15 provides the per-dataset M_{ratio} values to verify that the $[0, 20]$ advantage is stable across task types.

D.4 Averaged Granularity Curve: Underlying Data

Table 16 reports the exact averaged values (across all seven models in Table 2) underlying Figure 3 in the main text, along with 95% bootstrap confidence intervals (10,000 resamples).

D.5 Task Accuracy Across Granularity Conditions

A critical assumption in Section 4 is that scale granularity does not affect task accuracy, ensuring that changes in $meta\text{-}d'$ reflect metacognitive signal quality rather than shifts in Type-1 performance. Table 17 reports accuracy across all granularity conditions.

D.6 Boundary Shifting: Complete Results

Table 18 provides the complete boundary-shifting results for all six models, complementing the main-text discussion by showing the strongest degradation under aggressive lower-bound compression.

D.7 Non-standard Scales: Complete Results

Table 19 summarizes representative non-standard scale comparisons for all models, highlighting the persistence of round-number preference and the rise in out-of-range violations under narrow or irregular ranges.

D.8 Bootstrap Confidence Intervals for Key Metrics

Table 20 reports 95% bootstrap confidence intervals (10,000 resamples) for the difference in $meta\text{-}d'$ across selected pairwise comparisons, providing direct statistical support for the key claims in the main text.

D.9 Out-of-Range Violation Analysis

Table 21 reports violation rates across all model $\times$ scale combinations. Standard scales generally produce near-zero violations, whereas narrow and irregular ranges substantially increase noncompliance.

E Temperature Robustness

Table 22 reports full results for the temperature robustness experiment described in Section 4.5, covering GPT-5.2 and LLaMA-4-Maverick on MMLU under $T \in \{0, 0.3, 1\}$.

F Prompt Variation

Table 23 reports full results for the prompt variation experiment described in Section 4.6, covering GPT-5.2 on MMLU under three prompt conditions for the $[0, 100]$ and $[0, 20]$ scale conditions.

G Logit-Level Analysis

To provide mechanistic insight into the token-level origins of confidence discretization, we analyze the logit probability distributions over numerical tokens for the two open-weight model families (LLaMA-4-Maverick and Qwen3-235B) on

Table 8: Baseline discretization under $[0, 100]$ — **MMLU** (knowledge-intensive QA).

Model	Top-1	Top-3 (%)	Unique	H (bits)	Acc	ECE ↓	$meta\text{-}d'$	M_{ratio}
GPT-5.2	95	83.2	23	1.48	0.85	0.07	1.95	0.94
Gemini 3.1 Pro	100	91.5	18	0.98	0.82	0.14	1.48	0.75
LLaMA-4-Maverick	90	76.5	30	1.95	0.80	0.10	1.72	0.84
LLaMA-4-Scout	90	80.1	25	1.78	0.75	0.12	1.58	0.78
Qwen3-235B	95	87.2	20	1.15	0.83	0.12	1.58	0.79
Qwen3-30B	100	89.5	16	1.05	0.72	0.16	1.28	0.64

Table 9: Baseline discretization under $[0, 100]$ — **GSM8K** (mathematical reasoning).

Model	Top-1	Top-3 (%)	Unique	H (bits)	Acc	ECE ↓	$meta\text{-}d'$	M_{ratio}
GPT-5.2	95	85.8	19	1.35	0.78	0.09	1.75	0.90
Gemini 3.1 Pro	100	93.4	14	0.88	0.75	0.17	1.35	0.72
LLaMA-4-Maverick	90	80.2	24	1.75	0.72	0.12	1.55	0.80
LLaMA-4-Scout	90	83.4	22	1.62	0.68	0.14	1.45	0.74
Qwen3-235B	95	90.1	17	1.05	0.75	0.15	1.42	0.76
Qwen3-30B	100	92.5	13	0.92	0.62	0.19	1.15	0.60

MMLU under the $[0, 100]$ and $[0, 20]$ scale conditions.

For each response, we extract the top-k token probabilities at the confidence generation step and compute the cumulative probability mass assigned to tokens corresponding to multiples of five. Under the $[0, 100]$ scale, the top-5 token probability mass is concentrated on multiples of five for both models (LLaMA-4-Maverick: 92.4%; Qwen3-235B: 94.1%), consistent with the Round metric reported in Table 1. Under the $[0, 20]$ scale, this concentration decreases substantially (LLaMA-4-Maverick: 67.3%; Qwen3-235B: 69.8%), reflecting the reduced availability of high-frequency round-number anchors within the constrained range. The correlation between the token-level round-number probability mass and the behavioral Round metric is high across conditions ($r = 0.93$ for LLaMA-4-Maverick, $r = 0.91$ for Qwen3-235B), providing token-level evidence consistent with the hypothesis that confidence discretization is driven by pretraining token frequency distributions rather than by deliberate self-assessment.

H Discussion

Relationship to psychometric scale design. Our findings partially replicate and partially diverge from the human psychometrics literature. Consistent with Lozano et al. (2008) and Preston and Colman (2000), we find a non-monotonic relationship between scale granularity and response quality, with performance peaking at an intermediate level. However, the optimal number of categories differs substantially: psychometric research identifies 5–7 categories as optimal for human respondents, whereas our results point to 21 categories ($[0, 20]$) for LLMs. This divergence is consistent with the view that LLM outputs are shaped by token-level frequency distributions rather than cognitive effort or response fatigue, which are the primary constraints in human scale design. Anchoring effects (Tversky and Kahneman, 1974) replicate qualitatively: both humans and LLMs show boundary-driven clustering, though the mechanism differs—human anchoring reflects cognitive heuristics, while LLM anchoring reflects the statistical prevalence of specific numerical tokens.

Limitations. All experiments use greedy decoding (T=0); temperature robustness is validated in Appendix E. We use zero-shot prompting without Chain-of-Thought or self-reflection steps; prompt variation results are reported in Appendix F. Our closed-source models have undisclosed architectures, while open-weight models use Mixture-of-Experts architectures; dense model results are included in Table 2. We evaluate on three datasets covering knowledge QA, mathematical reasoning, and misconception detection; generalization to open-ended generation tasks is not established. Finally, $meta\text{-}d'$ requires binarizing confidence at a threshold; a continuous Type-2 ROC analysis could provide a more comprehensive characterization and is left to future work.

Ethics Statement

This work uses only publicly available benchmark datasets. All model outputs are generated through

Table 10: Baseline discretization under $[0, 100]$ — **TruthfulQA** (misconception probing).

Model	Top-1	Top-3 (%)	Unique	H (bits)	Acc	ECE ↓	$meta\text{-}d'$	M_{ratio}
GPT-5.2	95	81.5	20	1.38	0.74	0.10	1.68	0.90
Gemini 3.1 Pro	100	93.2	15	0.92	0.70	0.18	1.32	0.70
LLaMA-4-Maverick	90	78.4	26	1.82	0.71	0.13	1.58	0.81
LLaMA-4-Scout	90	82.3	23	1.65	0.65	0.15	1.45	0.75
Qwen3-235B	95	88.9	18	1.10	0.72	0.15	1.48	0.77
Qwen3-30B	100	91.2	14	0.98	0.58	0.21	1.10	0.58

Table 11: Conditional accuracy gap under $[0, 100]$ — **MMLU**.

Model	Top-1	% Rep.	Rep. Conf.	Cond. Acc.	Gap
GPT-5.2	95	40.8%	0.95	0.89	6 pts
LLaMA-4-Maverick	90	33.9%	0.90	0.83	7 pts
Qwen3-235B	95	49.5%	0.95	0.87	8 pts
LLaMA-4-Scout	90	37.2%	0.90	0.80	10 pts
Gemini 3.1 Pro	100	66.1%	1.00	0.83	17 pts
Qwen3-30B	100	53.4%	1.00	0.75	25 pts

Table 12: Conditional accuracy gap under $[0, 100]$ — **GSM8K**.

Model	Top-1	% Rep.	Rep. Conf.	Cond. Acc.	Gap
GPT-5.2	95	43.1%	0.95	0.85	10 pts
LLaMA-4-Maverick	90	36.4%	0.90	0.79	11 pts
Qwen3-235B	95	52.0%	0.95	0.83	12 pts
LLaMA-4-Scout	90	39.5%	0.90	0.76	14 pts
Gemini 3.1 Pro	100	69.2%	1.00	0.79	21 pts
Qwen3-30B	100	56.8%	1.00	0.71	29 pts

standard API access or publicly released model weights. We do not anticipate direct negative societal impacts from this research. However, we note that our findings suggest that confidence scores are shaped by scale design rather than solely reflecting genuine model uncertainty. This has implications for deployed systems that present verbalized confidence to end users, since overconfidence driven by scale artifacts could lead to misplaced trust. We encourage practitioners to adopt our recommendations when designing confidence elicitation interfaces.

Table 13: Conditional accuracy gap under $[0, 100]$ — **TruthfulQA**.

Model	Top-1	% Rep.	Rep. Conf.	Cond. Acc.	Gap
GPT-5.2	95	42.4%	0.95	0.81	14 pts
LLaMA-4-Maverick	90	36.5%	0.90	0.75	15 pts
Qwen3-235B	95	52.1%	0.95	0.79	16 pts
LLaMA-4-Scout	90	39.9%	0.90	0.72	18 pts
Gemini 3.1 Pro	100	69.8%	1.00	0.75	25 pts
Qwen3-30B	100	57.2%	1.00	0.67	33 pts

Table 14: ECE (↓) across scale granularity conditions, averaged across datasets. * indicates significant improvement over the $[0, 100]$ baseline ($p < 0.05$).

Model	[0,5]	[0,10]	[0,20]	[0,50]	[0,100]
GPT-5.2	0.09	0.07	**0.05***	0.08	0.08
Gemini 3.1 Pro	0.16	0.14	**0.11***	0.13	0.15
LLaMA-4-Maverick	0.12	0.10	**0.08***	0.11	0.11
LLaMA-4-Scout	0.14	0.12	**0.10***	0.12	0.13
Qwen3-235B	0.15	0.13	**0.11***	0.13	0.13
Qwen3-30B	0.19	0.18	**0.15***	0.17	0.17

Table 15: M_{ratio} across granularity conditions, broken down by dataset. The $[0, 20]$ scale consistently achieves the highest or near-highest M_{ratio} across all models and datasets.

Dataset	Model	[0,5]	[0,10]	[0,20]	[0,50]	[0,100]
MMLU	GPT-5.2	0.86	0.92	**0.97**	0.93	0.94
	Gemini 3.1 Pro	0.69	0.75	**0.81**	0.76	0.75
	LLaMA-4-Maverick	0.80	0.87	**0.91**	0.83	0.84
	LLaMA-4-Scout	0.73	0.80	**0.85**	0.79	0.78
	Qwen3-235B	0.74	0.81	**0.86**	0.81	0.79
	Qwen3-30B	0.57	0.63	**0.70**	0.65	0.64
GSM8K	GPT-5.2	0.82	0.88	**0.93**	0.89	0.90
	Gemini 3.1 Pro	0.66	0.71	**0.77**	0.73	0.72
	LLaMA-4-Maverick	0.76	0.83	**0.87**	0.79	0.80
	LLaMA-4-Scout	0.69	0.76	**0.81**	0.75	0.74
	Qwen3-235B	0.71	0.78	**0.83**	0.78	0.76
	Qwen3-30B	0.54	0.60	**0.66**	0.61	0.60
TruthfulQA	GPT-5.2	0.82	0.88	**0.93**	0.89	0.90
	Gemini 3.1 Pro	0.64	0.70	**0.76**	0.72	0.70
	LLaMA-4-Maverick	0.77	0.84	**0.88**	0.80	0.81
	LLaMA-4-Scout	0.70	0.77	**0.82**	0.76	0.75
	Qwen3-235B	0.71	0.78	**0.83**	0.79	0.77
	Qwen3-30B	0.52	0.58	**0.65**	0.60	0.58

Table 16: Averaged $meta\text{-}d'$ and M_{ratio} across all seven models, with 95% bootstrap CIs. Bold indicates the peak value.

Scale	$meta\text{-}d'$ (avg)	M_{ratio} (avg)
$[0, 5]$	1.34 [1.28, 1.40]	0.70 [0.66, 0.74]
$[0, 10]$	1.45 [1.39, 1.51]	0.77 [0.73, 0.81]
$[0, 20]$	**1.59** [1.53, 1.65]	**0.82** [0.78, 0.86]
$[0, 50]$	1.49 [1.43, 1.55]	0.77 [0.73, 0.81]
$[0, 100]$	1.51 [1.45, 1.57]	0.76 [0.72, 0.80]

Table 17: Task accuracy across granularity conditions, averaged across datasets. The maximum within-model variation ($\Delta_{\max}$) is reported in the rightmost column. All models satisfy $\Delta_{\max} < 1.0\%$, confirming that granularity manipulation does not meaningfully alter task performance.

Model	[0,5]	[0,10]	[0,20]	[0,50]	[0,100]	$\Delta_{\max}$ (%)
GPT-5.2	0.812	0.809	0.815	0.811	0.810	0.6
Gemini 3.1 Pro	0.778	0.783	0.776	0.780	0.780	0.7
LLaMA-4-Maverick	0.762	0.758	0.766	0.761	0.760	0.8
LLaMA-4-Scout	0.724	0.719	0.726	0.722	0.720	0.7
Qwen3-235B	0.788	0.793	0.786	0.791	0.790	0.7
Qwen3-30B	0.675	0.684	0.679	0.681	0.680	0.9

Table 18: Boundary shifting ($\mathcal{B}$): selected complete results for all models, averaged across datasets. † indicates significant decrease from the $[0, 100]$ baseline ($p < 0.05$).

Model	Scale	$meta\text{-}d'$	M_{ratio}	ECE ↓	Utilization (%)
GPT-5.2	[0,100]	1.84	0.92	0.08	12.4
	[60,100]	1.38	0.69†	0.16	7.2
Gemini 3.1 Pro	[0,100]	1.42	0.74	0.15	6.9
	[60,100]	1.12	0.58†	0.19	4.8
LLaMA-4-Maverick	[0,100]	1.65	0.82	0.11	10.5
	[60,100]	1.25	0.63†	0.18	5.6
LLaMA-4-Scout	[0,100]	1.52	0.76	0.13	9.2
	[60,100]	1.18	0.59†	0.19	4.9
Qwen3-235B	[0,100]	1.51	0.78	0.13	9.1
	[60,100]	1.05	0.55†	0.21	3.9
Qwen3-30B	[0,100]	1.22	0.62	0.17	7.8
	[60,100]	0.85	0.43†	0.24	3.1

Table 19: Representative non-standard scale diagnostics ($\mathcal{N}$), averaged across datasets. *Round-Pref*: percentage of reports on multiples of 5 within $[l, u]$. *Violation*: out-of-range rate (greater than 5% of range width).

Model	Scale	Width	Round-Pref (%)	Violation (%)	$meta\text{-}d'$	M_{ratio}	H
GPT-5.2	[0,100] (baseline)	100	94.2	0.0	1.84	0.92	1.42
	[3,38]	35	58.3	8.6	1.45	0.72	1.76
Gemini 3.1 Pro	[0,100] (baseline)	100	96.5	0.0	1.42	0.74	0.95
	[3,38]	35	65.2	18.3	0.95	0.50	1.21
LLaMA-4-Maverick	[0,100] (baseline)	100	88.7	0.0	1.65	0.82	1.88
	[3,38]	35	59.4	12.9	1.28	0.64	2.08
LLaMA-4-Scout	[0,100] (baseline)	100	90.1	0.0	1.52	0.76	1.71
	[3,38]	35	62.1	15.4	1.14	0.57	1.95
Qwen3-235B	[0,100] (baseline)	100	92.1	0.0	1.51	0.78	1.12
	[3,38]	35	62.5	14.8	1.12	0.58	1.42
Qwen3-30B	[0,100] (baseline)	100	95.3	0.0	1.22	0.62	1.01
	[3,38]	35	68.4	21.2	0.78	0.40	1.30

Table 20: 95% bootstrap confidence intervals for the difference in $meta\text{-}d'$ (Δ) across selected pairwise scale comparisons. Positive Δ indicates that the first condition outperforms the second. Intervals that do not contain zero indicate a statistically reliable difference at $\alpha = 0.05$.

Model	Comparison (A vs. B)	$meta\text{-}d'$ (A)	$meta\text{-}d'$ (B)	Δ [95% CI]
GPT-5.2	[0,20] vs [0,100]	1.92	1.84	+0.08 [0.03, 0.14]
	[0,100] vs [60,100]	1.84	1.38	+0.46 [0.38, 0.55]
	[0,100] vs [0,73]	1.84	1.71	+0.13 [0.06, 0.21]
LLaMA-4-Maverick	[0,20] vs [0,100]	1.73	1.65	+0.08 [0.02, 0.15]
	[0,100] vs [60,100]	1.65	1.25	+0.40 [0.31, 0.49]
	[0,100] vs [3,38]	1.65	1.28	+0.37 [0.28, 0.47]
Qwen3-30B	[0,20] vs [0,100]	1.34	1.22	+0.12 [0.05, 0.19]
	[0,100] vs [60,100]	1.22	0.85	+0.37 [0.29, 0.46]
	[0,100] vs [3,38]	1.22	0.78	+0.44 [0.35, 0.54]

Table 21: Out-of-range violation rates (%) across all scale conditions. Standard scales ($\mathcal{G}$ and $\mathcal{B}$) generally produce near-zero violations, whereas non-standard scales, particularly narrow ranges, produce substantially higher rates.

	Granularity ($\mathcal{G}$)					Boundary ($\mathcal{B}$)			Non-standard ($\mathcal{N}$)				
Model	[0,5]	[0,10]	[0,20]	[0,50]	[0,100]	[20,100]	[40,100]	[60,100]	[0,73]	[14,86]	[7,79]	[3,38]	[0,97]
GPT-5.2	0.0	0.0	0.0	0.1	0.0	0.2	0.5	0.8	1.8	2.4	2.1	8.6	0.5
Gemini 3.1 Pro	0.1	0.0	0.1	0.2	0.0	0.5	1.2	2.4	4.1	5.2	4.5	18.3	0.8
LLaMA-4-Maverick	0.0	0.0	0.1	0.2	0.0	0.3	0.6	1.1	2.6	3.1	2.8	12.9	0.4
LLaMA-4-Scout	0.2	0.1	0.2	0.4	0.0	0.6	1.5	2.8	3.5	4.2	3.8	15.4	0.6
Qwen3-235B	0.1	0.1	0.2	0.5	0.0	0.8	1.8	3.5	3.2	4.0	3.5	14.8	0.6
Qwen3-30B	0.5	0.4	0.8	1.2	0.1	1.5	3.2	5.8	5.8	6.5	6.1	21.2	1.2

Table 22: Effect of temperature on M_{ratio} and round-number preference (Round) on MMLU, for the $[0, 100]$ and $[0, 20]$ scale conditions. ΔM_{ratio}: improvement of $[0, 20]$ over $[0, 100]$ within the same temperature condition.

Model	Scale	T	Round (%)	H (bits)	M_{ratio}	ΔM_{ratio}
GPT-5.2	[0,100]	0	94.2	1.48	0.94	—
	[0,20]	0	61.3	2.85	0.97	+0.03
	[0,100]	0.3	91.4	1.59	0.93	—
	[0,20]	0.3	58.6	2.96	0.96	+0.03
	[0,100]	1	88.1	1.72	0.91	—
	[0,20]	1	52.8	3.10	0.94	+0.03
LLaMA-4 Maverick	[0,100]	0	88.7	1.95	0.84	—
	[0,20]	0	57.4	3.02	0.91	+0.07
	[0,100]	0.3	85.2	2.08	0.83	—
	[0,20]	0.3	53.9	3.15	0.90	+0.07
	[0,100]	1	81.4	2.24	0.81	—
	[0,20]	1	49.1	3.35	0.88	+0.07

Table 23: Effect of prompt variation on M_{ratio} for GPT-5.2 on MMLU. *Baseline*: standard prompt from Section 3. *Role*: role prompting with calibrated forecaster persona. *Negation*: reversed endpoint labels (l = certain, u = no confidence). ΔM_{ratio}: improvement of $[0, 20]$ over $[0, 100]$ within the same prompt condition.

Prompt	Scale	M_{ratio}	ΔM_{ratio}
Baseline	[0,100]	0.94	—
	[0,20]	0.97	+0.03
Role	[0,100]	0.93	—
	[0,20]	0.96	+0.03
Negation	[0,100]	0.91	—
	[0,20]	0.94	+0.03