Asking Is Not Enough: Protocol Sensitivity in LLM Confidence Calibration

Anonymous ACL submission

Abstract

LLM confidence calibration is often evaluated by comparing two signals: token-probability scores and verbalized confidence. These signals are sometimes treated as direct readouts of model uncertainty, but their comparison depends on measurement choices that are rarely made explicit. In the main analysis, we hold the verbalized-confidence elicitation fixed: a single prompt template, probability scale, and output format. We then vary the measurement axes that define the verbalized-vs-token comparison: which answer string receives the token-probability score, how that score is read from the answer tokens, and under which conditioning context it is measured. We evaluate this design on four QA benchmarks across three open 7–8B base/Instruct model families, with larger Qwen2.5 variants as same-family robustness checks. The resulting comparison is sensitive to these choices: conditioning context changes the sign or magnitude of the ECE gap across settings, token readout produces smaller but still sign-moving changes, and changing the ECE estimator has little effect. Under the default generated-answer, bare-context protocol, Instruct settings are close to parity rather than showing a large calibration gain for verbalized confidence. In a separate supplied-answer analysis, surface-plausible wrong answers receive nearly the same confidence as supplied gold answers, suggesting that verbalized confidence also reflects answer plausibility and provenance rather than correctness alone. We argue that both confidence signals should be treated as protocol-dependent behavioral measurements, and provide a reporting checklist covering elicitation provenance, scored answer, token-probability readout, and conditioning context.

1 Introduction

Calibration of language model confidence underpins safe deployment in high-stakes settings, from medical question answering to coding agents. Recent work therefore evaluates not only whether a model answers correctly, but also whether its stated confidence is meaningful. Two confidence signals are especially common. One is read from the model's token probabilities; the other is *verbalized confidence*, where the model is asked to report a probability in text (Tian et al., 2023; Kadavath et al., 2022; Xiong et al., 2024; Yang et al., 2024; Dai, 2026). Verbalized confidence is attractive because it is easy to elicit from closed or chat-only systems, but it also turns confidence estimation into an interaction protocol: the answer may be self-generated or supplied, the confidence may be requested in different formats, and the token probability used for comparison must be read from some context and some position.

Prior work has shown that verbalized confidence is sensitive to how it is elicited (Xiong et al., 2024; Yang et al., 2024; Dai, 2026). We study a complementary source of sensitivity. Even when the prompt template is held fixed, the comparison between a verbalized self-report and a token-probability score can change with token-side and answer-provenance choices that are usually left implicit. Such choices matter because they define the quantity being measured: whether the confidence is attached to the model's own answer or a supplied answer, whether the token probability is read under the bare query or after a verbalized prefix, and how answer-token probabilities are read or aggregated. If these choices are not reported, a scalar calibration improvement can be mistaken for a property of the model or alignment procedure rather than a property of the measurement protocol.

We examine this issue through a controlled case study of the self-report/token-probability comparison. Whenever measuring verbalized confidence, our main analysis uses the same continuous 0.0–1.0 prompt template and evaluates three open base/Instruct families (Llama-3-8B, Mistral-

7B-v0.3, and Qwen-2.5-7B) on TriviaQA, SciQ, TruthfulQA, and MMLU, with larger-model robustness checks within the Qwen2.5 family. We report the gap $g = \mathrm{ECE}(c_{\mathrm{verb}}) - \mathrm{ECE}(c_{\mathrm{tok}})$ under alternative token-probability and answer choices, using the standard 10-bin ECE and bin-free calibration estimators. The goal is not to decide which confidence signal is universally correct; it is to test whether the conclusion of the comparison is stable once the protocol is made explicit.

The comparison is not stable under these choices. Token readout and conditioning context each change the sign or magnitude of the observed gap across many model–dataset settings, with context shifts comparable to changing the evaluation dataset; by contrast, changing the ECE estimator has little effect in our experiments. The results also do not support a large, protocol-independent calibration gain for verbalized confidence: under the specified generated-answer, bare-context protocol, most Instruct settings sit near parity rather than showing a large ECE reduction. Finally, when the model is asked to assign confidence to a supplied answer, the reported number distinguishes clearly mismatched answers from plausible ones but barely separates plausible wrong answers from correct ones. This suggests that verbalized confidence can track answer plausibility and provenance rather than correctness alone.

These findings motivate a more conservative interpretation of LLM confidence measurements. Neither verbalized confidence nor token probability should be treated as a direct readout of model uncertainty; both are behavioral measurements whose interpretation depends on how the answer and probability are elicited, conditioned, and scored. Claims that alignment improves self-reported uncertainty, or that verbalized confidence should be preferred to token probabilities, therefore require the measurement protocol to be specified first: which answer the confidence attaches to, under which context the token probability is measured, and how the token-probability score is read from the answer. Because different protocol choices define different measured quantities, we interpret our comparisons as sensitivity analyses rather than as a causal decomposition.

The contributions of this paper are as follows:

- **Protocol sensitivity with fixed verbalized-confidence elicitation.** We use the same prompt template, confidence scale, and output format whenever measuring verbalized confidence, and vary the answer and token-probability choices: answer provenance, token-probability readout, conditioning context, and estimator choice. Token readout and conditioning context materially change the observed ECE gap, whereas changing the ECE estimator has little effect; the context effect is comparable to changing the model family or dataset (§4.1).
- **Answer provenance and plausibility.** We compare verbalized confidence on self-generated answers with confidence elicited on supplied correct answers and supplied plausible wrong answers. The reported number barely separates a plausible wrong answer from a correct one, even though it drops for clearly mismatched answers, showing that verbalized confidence is not simply a correctness estimate (§4.2).
- **Reporting checklist and robustness.** We distill the results into a reporting checklist for confidence-signal comparisons: elicitation provenance, scored answer, token-probability readout, and conditioning context (§4.3). The conditioning-context sensitivity persists under an integer 0–100 prompt scale, four non-ECE metrics, and larger-model robustness checks within the Qwen2.5 family, while the finding that estimator choice has little effect is stable to bin count (Appendices I, J, K, N, L).

2 Related Work

Verbalized confidence. LLM confidence is commonly evaluated either from token probabilities or from confidence values stated in text. Earlier work studies natural-language uncertainty more broadly, including calibrated uncertainty expressions in words (Lin et al., 2022a) and linguistic calibration in dialogue (Mielke et al., 2022). Tian et al. (2023) ask RLHF-tuned chat models to output continuous probabilities and compare them with token-probability scores, while Kadavath et al. (2022) elicit a "P(True)" judgment on a supplied candidate answer. Later work shows that verbalized confidence is sensitive to how it is elicited: Xiong et al. (2024) find systematic overconfidence and mixed benefits from prompting or aggregation strategies, Yang et al. (2024) compare many prompt variants and find strong model and dataset dependence, and Dai (2026) shows that the response scale can shape

the reported numbers. This literature establishes verbalized confidence as an elicited behavior rather than a context-free uncertainty readout. We ask a complementary question: when the verbalized-confidence measurement uses a common format, probability scale, and output convention, is its comparison with token-probability scores stable under choices about the answer being scored and the token-probability readout?

Confidence signals and calibration metrics. Token probabilities are also not protocol-free: an answer-level score requires choosing the conditioning context, the token span or position, and the correctness label. Prior NLP work studies whether pretrained transformer probabilities are calibrated across tasks and distribution shifts (Desai and Durrett, 2020). Other work proposes alternative confidence or uncertainty signals, including Rank-Calibration (Huang et al., 2024), sample consistency (Lyu et al., 2024), and representation stability (Khanmohammadi et al., 2025). A separate calibration-estimation literature shows that ECE itself depends on binning and estimator choices (Guo et al., 2017; Nixon et al., 2019; Kumar et al., 2019; Roelofs et al., 2022; Zhang et al., 2020; Błasiok and Nakkiran, 2024; Gupta and Ramdas, 2021). This motivates checking whether a confidence comparison is robust to reasonable estimator or metric choices, while keeping those choices distinct from protocol choices about which token probability is read.

Protocol specification. Prompt-format sensitivity is a known issue in few-shot LM evaluation and motivates contextual calibration (Zhao et al., 2021). Surveys of LLM confidence estimation identify prompt sensitivity, metric choice, and confidence-signal choice as recurring limitations (Geng et al., 2024), and empirical work shows that verbalized confidence can change with elicitation and scale design (Xiong et al., 2024; Yang et al., 2024; Dai, 2026). We focus on protocol choices that define the comparison itself: elicitation provenance, scored answer, token position, and conditioning context. A confidence value may refer to a self-generated answer or to a supplied candidate answer; a token-probability score may be read under the bare query or after the prompt prefix used to elicit verbalized confidence; and different token positions can correspond to different measured quantities. Our contribution is a sensitivity analysis and reporting checklist for these choices, so that calibration differences are interpreted as protocol-dependent measurements rather than direct properties of a model or alignment procedure.

3 Setup and Method

3.1 Models, Data, and Confidence Measures

We evaluate three open base/Instruct pairs: Llama-3-8B (Meta AI, 2024), Mistral-7B-v0.3 (Jiang et al., 2023), and Qwen2.5-7B (Qwen Team, 2024), with their Instruct variants. We use four QA benchmarks: TriviaQA (Joshi et al., 2017) (1k subset), SciQ (Welbl et al., 2017) (1k), TruthfulQA (Lin et al., 2022b) (817 full), and four MMLU (Hendrycks et al., 2021) subjects (high-school physics, philosophy, high-school government and politics, and miscellaneous; 1,155 examples, per-subject sizes capped by availability). The verbalized-confidence measurement uses one continuous 0.0–1.0 probability template throughout, applied through each Instruct model's chat template and as raw text for base models.

We vary answer-side and token-probability choices to probe what the confidence measures track, and to quantify how protocol choices affect the verbalized-vs-token comparison:

- **Answer provenance for verbalized confidence.** Verbalized confidence is typically elicited by asking the model to answer and then report confidence in that answer. If this measure were provenance-invariant, correct self-generated answers and supplied gold answers should receive similar confidence even when they are not identical strings; exact string matches are analyzed separately as a matched-string control. Supplied wrong answers should receive lower confidence. We therefore compare the model's own answer with three prefilled-answer conditions, where the answer string is inserted into the confidence prompt and the model generates only the confidence value: the gold answer, a same-question plausible wrong answer generated by another model, and an off-type wrong answer taken from a different question. This separates correctness sensitivity from answer source and plausibility.
- **Scored answer for token probability.** The same answer-identity issue also affects token-probability scoring: the model-generated answer and the dataset's gold reference can be

Axis	Main setting	Alternative/check	Held fixed	Contrast
Verb. answer provenance	self-generated	supplied gold/plaus./off-type	prompt, scale, models, data	supplied-answer
Token-prob. scored answer	model-generated	gold ref. (TF)	verb. values, labels, context, T_1, ECE	generated-vs-gold
Token readout	answer-span mean (T_1)	first answer token (T_2)	scored answer, context, ECE	readout shift
Conditioning context	bare query	verb. prefix	scored answer, T_1, ECE	g_A vs. g_B
Estimator	10-bin ECE	bin-free mean	scored answer, context, T_1	g_A vs. g_C

Table 1: Protocol axes and contrasts. The main ECE setting is the generated-answer, bare-query, T_1, 10-bin protocol. "TF" denotes teacher forcing on the gold reference. The supplied-answer analysis elicits verbalized confidence only; the remaining rows define verbalized-vs-token ECE contrasts. The T_1 and T_2 token readouts are defined in the text. The A–D notation is used only for the conditioning-context/estimator block reported in Table 2.

different strings even when both are correct. For the ECE comparison, the default token-probability score is assigned to the model's freely generated answer. As a teacher-forced check, we instead append the gold string to the bare query and read its conditional probability in a single forward pass, while holding the verbalized-confidence values and correctness labels fixed to the model-generated answer. We refer to this contrast as *generated-vs-gold*.

- **Token-probability readout.** Because a generated answer can span multiple tokens $a_1, \ldots, a_L$, a token-probability score must specify whether it summarizes the whole answer span or only the initial answer token. We use either the answer-span geometric mean $T_1 = \exp\left(\frac{1}{L}\sum_{i=1}^{L} \log p(a_i \mid h, a_{<i})\right)$ or the first answer-token probability $T_2 = p(a_1 \mid h)$, where h is the conditioning context; the two coincide only for one-token answers.

- **Conditioning context.** Verbalized confidence requires a confidence-elicitation prefix that asks the model to report a probability, whereas token probability can be measured from the bare query alone. Computing the two signals under these different contexts introduces a context mismatch; computing token probability under the same verbalized prefix removes that mismatch but changes the conditional distribution being read. We therefore measure token probability either under the bare query or under the same verbalized prompt prefix that elicits the model's confidence.

- **Calibration estimator.** Computing ECE requires specifying an estimator. Our default is equal-width 10-bin ECE, which partitions confidence values into ten intervals on $[0, 1]$. As a bin-free check, we also compute KDE-ECE (Zhang et al., 2020), KS-Cal (Gupta and Ramdas, 2021), and Smoothed ECE (Błasiok and Nakkiran, 2024), and summarize them by their mean. This summary is used only to check whether the result changes outside the equal-width binning scheme; we do not propose it as a new calibration metric.

Table 1 summarises these axes together with the alternative, the choices held fixed, and the contrast reported under each.

The default ECE protocol assigns token probability to the model-generated answer, reads it under the bare query using T_1, and computes 10-bin ECE. We then vary these axes to test how sensitive the confidence measures and their comparison are to measurement choices, while keeping the confidence scale and reported-value format fixed. In the supplied-answer analysis, we keep the confidence prompt fixed and vary only the answer string placed in the `Guess:` slot.

In the main protocol, for each (model, dataset, prompt-mode) setting, we decode greedily ($T = 0$), parse the generated answer, score it for correctness, and then compute token probability on that generated answer. We parse verbalized confidence with a fixed, pre-specified parser. Rows without parseable confidence are excluded from verbalized-confidence metrics, whereas token-probability metrics are computed on the full sample. This creates only a small sample mismatch for the Instruct models used in our main analyses, where parse rates are at least 0.966 (Appendix E); for base models, we also report a `qid`-matched robustness check in Appendix F.

Answers are scored with one lenient string-

matching rule across datasets. Before comparison, predictions and gold strings are lowercased, non-alphanumeric characters are replaced with spaces, and multiple spaces are collapsed into one; a prediction is correct if the normalized strings match exactly or one contains the other. For bare-query outputs, we take the first generated line after removing common answer prefixes. For verbalized-confidence outputs, we take the answer string after `Guess:`. For SciQ and MMLU, we use the datasets as free-form QA sources rather than letter-choice tasks: models generate answer strings, and gold/supplied answers use the text of the correct option rather than its option letter. Thus, in generated-vs-gold and supplied-answer checks on SciQ and MMLU, the comparison is between the model's generated answer string and the correct option text, not between multiple-choice labels. TriviaQA uses aliases; TruthfulQA uses the reference correct-answer set. Full dataset-specific details are in Appendix M.

3.2 Protocol Choices and Metrics

For each measurement configuration X for which both confidence signals are defined, we report the signed calibration gap

$$g_X = \mathrm{ECE}(c_{\text{verb},X}, y) - \mathrm{ECE}(c_{\text{tok},X}, y),$$

where $c_{\text{verb},X}$ is the parsed verbalized confidence under configuration X, $c_{\text{tok},X}$ is the corresponding token-probability score, y is the binary correctness label, and ECE is computed using the estimator specified by X. In the main ECE contrasts, y is held fixed to the correctness of the model-generated answer unless stated otherwise. Negative values mean verbalized confidence has lower ECE. For the equal-width binned estimator, ECE averages the absolute gap between mean confidence and empirical accuracy within confidence bins, weighted by the number of examples in each bin. The main configuration, g_A, is the generated-answer, bare-query, T_1, 10-bin protocol.

The A–D notation is used for the context and estimator variants: **A**, bare-query token probability with 10-bin ECE; **B**, token probability measured after the verbalized prompt prefix with 10-bin ECE; **C**, bare-query token probability with the bin-free estimator mean (KDE-ECE (Zhang et al., 2020), KS-Cal (Gupta and Ramdas, 2021), and Smoothed ECE (Błasiok and Nakkiran, 2024)); and **D**, both changes applied. We report each g_X with a 95% bootstrap CI and report pairwise differences such as $g_A - g_B$ as descriptive changes in the observed gap. Other axes in Table 1, including token readout, scored answer, and supplied-answer provenance, are reported separately as targeted contrasts rather than as part of the A–D grid.

Configuration B measures token probability under the same prefix used for verbalized-confidence elicitation. This prefix asks the model to complete a guess-and-probability format, so the resulting score is an operationally different conditional probability from the one read under the bare-query prompt in A. Appendix C shows that this context shift is not driven by the token-probability readout: the shift has the same sign under both the answer-span score T_1 and the first-token score T_2 on all Instruct settings (12/12).

We include two auxiliary diagnostics to interpret the main protocol effects. For the conditioning-context change, we report first-step next-token KL, final-layer CKA distance (Kornblith et al., 2019), and answer-change rate between the bare and verbalized-prefixed contexts (Appendix B). For verbalized outputs, we measure concentration near the $\{0.0, 0.1, \ldots, 1.0\}$ grid to assess whether round-number clustering could account for estimator-dependent patterns.

All ECEs, gaps, and pairwise differences are reported with non-parametric bootstrap 95% confidence intervals using $B = 1000$ resamples. Cross-setting aggregates use macro means with macro-bootstrap intervals. Appendix diagnostic tables use $B = 200$ resamples and are presented as descriptive robustness checks.

3.3 Supplied-Answer Analysis

The verbalized prompt asks for "the probability that the guess is correct." To test whether verbalized confidence tracks the correctness of the answer itself, or also depends on answer source and surface plausibility, we compare verbalized confidence across self-generated and supplied-answer provenance conditions. The self-generated condition corresponds to the main protocol, in which the model first generates an answer, and confidence is then elicited for that answer. For the three supplied-answer conditions, we use prefilled-answer elicitation: an answer string is placed in the `Guess:` slot, no answer tokens are sampled, and the model generates only the `Probability:` value. Because these conditions elicit $c_{\text{verb},X}$ but do not define a matched $c_{\text{tok},X}$, they are reported as provenance

Quantity	Instruct (n=12)	Base (n=12)
g_A (generated-answer, bare context)	-0.029 $[-0.042, -0.017]$	$+0.221$ $[+0.205, +0.238]$
g_B (verbalized-prefix context)	$+0.060$ $[+0.046, +0.074]$	$+0.346$ $[+0.330, +0.363]$
g_C (bin-free estimator)	-0.028 $[-0.041, -0.016]$	$+0.208$ $[+0.192, +0.225]$
g_D (both applied)	$+0.069$ $[+0.055, +0.082]$	$+0.336$ $[+0.320, +0.353]$
Token readout ($g_A^{T_1} - g_A^{T_2}$)	-0.031 $[-0.066, +0.005]^{\text{ns}}$	-0.082 $[-0.116, -0.049]$
Conditioning context ($g_A - g_B$)	-0.090 $[-0.103, -0.077]$	-0.125 $[-0.135, -0.115]$
Estimator choice ($g_A - g_C$)	-0.001 $[-0.003, +0.001]^{\text{ns}}$	$+0.013$ $[+0.009, +0.018]$
Second-order interaction	$+0.007$ $[+0.005, +0.009]$	$+0.003$ $[+0.000, +0.005]$

Table 2: Macro signed gaps and protocol shifts, averaged over twelve model-dataset settings per variant group, with 95% bootstrap CIs. Negative g means verbalized confidence has lower ECE. ^{ns} indicates that the macro CI overlaps zero. The verbalized-prefix context matches the prefix used to elicit verbalized confidence, but it also changes the conditional distribution being measured. Conditioning context moves the observed gap much more than estimator choice; token readout shows smaller but sign-moving cell-level sensitivity.

Setting	$\bar{g}_A$	$\overline{\lvert g_A - g_B \rvert}$	$\overline{\lvert g_A - g_C \rvert}$	flips
7–8B Instr.	-0.029	0.350	0.009	9/12
Qwen2.5-14B	-0.098	0.369	0.002	2/4
Qwen2.5-32B	-0.095	0.382	0.002	2/4
Qwen2.5-72B	-0.129	0.413	0.004	2/4

Table 3: Same-family larger-model robustness summary. Columns report the mean signed main-protocol gap, the mean absolute conditioning-context shift, the mean absolute estimator shift, and the number of datasets/settings whose g sign flips between g_A and g_B.

Source of variation	mean $\lvert\Delta g\rvert$
Variant (base vs. Instruct)	0.41
Model family	0.37
Conditioning context (protocol)	**0.28**
Dataset identity	0.17
Token readout (protocol)	**0.11**
ECE estimator (protocol)	**0.01**

Table 4: Mean absolute shift in g when one factor is changed. Protocol axes use Table 2; identity axes regroup g_A over 24 settings.

diagnostics rather than as g_X gaps. We compare four answer sources:

- **self**: the model's own guess from the main protocol.
- **gold**: the dataset's gold answer, prefilled.
- **plausible wrong**: another model's incorrect guess for the same question, prefilled.
- **off-type wrong**: a mismatched gold answer from another question, prefilled when no cross-model wrong answer is available.

In the main supplied-answer analysis, across the twelve 7–8B Instruct model-dataset settings, 7,229 of 11,916 supplied wrong-answer cases (60.7%; range across settings: 39.5–78.5%) use cross-model plausible wrong answers; the remaining 4,687 use off-type fallbacks. Appendix H reports the construction counts together with the per-setting provenance results. The supplied-answer conditions follow the Kadavath et al. (2022) "P(True)" style of elicitation, whereas the self-generated condition follows the main verbalized-confidence setup. For the main analysis, we run all three 7–8B Instruct models on all four datasets with greedy decoding; the `Probability:` value is recovered for 100% of items in every setting. The results also include a same-family Qwen2.5-14B-Instruct supplied-answer check.

4 Results

4.1 The Confidence Gap Is Protocol-Sensitive

Table 2 summarizes the main protocol-sensitivity results as macro means over twelve model-dataset settings per variant group; the underlying per-setting values are reported in Appendix A. We begin with the default protocol: verbalized confidence from the continuous probability template, token probability assigned to the model-generated answer, the bare-query T_1 answer-span readout, and 10-bin ECE. We report the signed gap $g = \mathrm{ECE}(c_{\text{verb}}) - \mathrm{ECE}(c_{\text{tok}})$, so negative values mean verbalized confidence has lower ECE. Under this protocol, Instruct models are close to parity rather than showing a large calibration gain for verbalized confidence. Compared with the large ECE reductions reported by Tian et al. (2023) for RLHF-tuned chat models, our open Instruct settings are much closer to parity under the default protocol. Averaged over the twelve Instruct model-dataset settings, the signed gap is $g_A = -0.029$ (95%

CI $[-0.042, -0.017]$): verbalized confidence is slightly better on average, but the effect is small and only 7 of 12 Instruct settings have $g_A < 0$. Base-model rows are included as secondary stress tests because parseable-confidence rates are uneven for some base models (Appendix E).

This near-parity conclusion is sensitive to how token probability is read. **Conditioning context** is the largest tested protocol axis: measuring token probability after the verbalized prompt prefix rather than under the bare query shifts the signed macro gap by -0.090 on Instruct and -0.125 on base, with multiple sign flips. **Token readout** is weaker but non-negligible: its Instruct macro CI overlaps zero, but it shifts the macro gap by -0.082 on base and changes the sign across T_1 and T_2 on 6 of 24 settings. By contrast, **estimator choice** has little effect: replacing 10-bin ECE with the bin-free estimator mean shifts the gap by only -0.001 on Instruct and $+0.013$ on base, and the bin-count sweep in Appendix D leaves this conclusion unchanged.

Table 3 extends the analysis to same-family Qwen2.5-Instruct models at 14B, 32B, and 72B. These runs are intended to assess whether the 7–8B pattern is specific to the main model scale, rather than to characterize scaling trends. The qualitative pattern is unchanged: the conditioning-context shift remains much larger than the estimator shift, and it reverses the sign of g on two of four datasets for each larger model. Because context shifts can have different signs across datasets, signed averages can hide large setting-level changes. Table 3 therefore reports mean absolute shifts across settings; the underlying per-setting values are given in Appendices A, J, K, and N.

Protocol choices shift g on a scale comparable to model and dataset variation. The conditioning-context axis has a mean absolute shift of 0.28, larger than the dataset-identity shift (0.17) and comparable to model family (0.37) and base/Instruct variant (0.41). Token readout (0.11) is also close to dataset identity, whereas estimator choice (0.01) is much smaller (Table 4).

Appendix D shows that the small estimator effect is not specific to the 10-bin choice. Appendix B reports auxiliary context-shift measurements, which describe how the prompt change affects model behavior but are not used as causal explanations for the ECE shift.

Variant	macro g_{gen}	macro g_{gold}	sign-flip
Instruct (n=12)	-0.029	$+0.083$	9/12
Base (n=12)	$+0.223$	$+0.288$	4/12
Qwen2.5-14B (n=4)	-0.098	$+0.000$	2/4

Table 5: Verbalized-vs-token gap under generated-answer (g_{gen}, main) vs. gold-teacher-forced (g_{gold}) token-probability scoring, with the correctness label held fixed. The scored-answer mismatch reverses the macro sign on the Instruct settings and on $2/4$ Qwen2.5-14B datasets; the base regime is secondary (parse-rate caveat, Appendix E).

4.2 Answer Choice and Provenance Affect Confidence Comparisons

The main protocol assigns token probability to the model-generated answer. We next compare it with a teacher-forced gold-answer scoring variant, holding correctness labels fixed to the model-generated answer. This changes the measured quantity: on Instruct settings, the macro gap moves from -0.029 under the generated-answer protocol to $+0.083$ under the gold-answer protocol, reversing sign on 9 of 12 settings (Table 5). The token-probability side is therefore sensitive to which answer string is scored, even when the verbalized-confidence values and correctness labels are held fixed. Appendix G explains this scoring contrast and visualizes the per-setting sign changes.

Table 6 reports verbalized confidence when the `Guess:` slot is filled with the model's own answer, a supplied gold answer, or supplied wrong answers (§3.3). The key pattern is that supplied surface-plausible wrong answers receive nearly the same confidence as supplied gold answers. Their marginal means differ by only 0.049 (0.783 vs. 0.832); within the same question, supplied gold answers receive only $+0.021$ higher confidence than supplied plausible wrong answers on average (95% CI $[+0.014, +0.028]$). Off-type wrong answers receive much lower confidence (0.240), indicating that verbalized confidence is sensitive to the supplied answer content and does not reflect correctness alone.

Table 6 also shows a modest preference for self-generated answers: confidence is higher for self-generated correct answers than for supplied gold answers by $+0.076$, and higher for self-generated wrong answers than for supplied surface-plausible wrong answers by $+0.024$. The preference is therefore larger when the model's own answer is correct. For correct answers, we fur-

Quantity	value
self correct mean	0.908
self wrong mean	0.807
supplied gold mean	0.832
supplied plaus. wrong mean	0.783
supplied off-type mean	0.240
marginal gold − plaus. wrong	+0.049
paired gold − plaus. wrong	+0.021
self − gold, diff. strings	+0.055
self − gold, exact match	+0.000
gold − self, initially wrong	−0.010
correct diff.-string share	55.6%

Table 6: Verbalized confidence Y by `Guess:`-slot fill (3 Instruct models $\times$ 4 datasets, $n = 11{,}916$). Mean rows report Y; contrast rows report marginal or paired mean differences. Bootstrap CIs for paired contrasts are reported in Table 14.

ther split the paired comparison by whether the model's answer exactly matches the supplied gold string. When both are correct but the strings differ (55.6% of correct items), the self-generated answer receives $+0.055$ higher confidence than the supplied gold (95% CI $[+0.046, +0.064]$). When the strings exactly match, the difference is essentially zero. For questions the model initially answers incorrectly, supplied gold answers do not receive higher confidence than the model's own wrong answers on average (-0.010, 95% CI $[-0.016, -0.004]$). Same-family Qwen2.5-14B-Instruct results show the same supplied-answer pattern: supplied gold, surface-plausible wrong, and off-type wrong answers receive macro mean confidences 0.799, 0.705, and 0.071, respectively; the pooled differing-string self-preference is also positive ($+0.104$; Table 15). This single-model result is reported for robustness rather than to estimate scaling trends.

4.3 Reporting Implications

A verbalized-vs-token confidence comparison is not fully interpretable without its measurement configuration. Our results support reporting four measurement axes:

C1 Elicitation provenance: whether verbalized confidence is elicited for the model's self-generated answer or, in a supplied-answer diagnostic, for an answer supplied in the prompt.

C2 Scored answer: which answer string receives the token-probability score, such as the model-generated answer or a teacher-forced gold reference.

C3 Token-probability readout: for example, the answer-span geometric mean (T_1) versus the first answer-token probability (T_2).

C4 Conditioning context: bare query versus the verbalized prompt-template prefix for the token-probability measurement.

The ECE estimator should also be specified, but in our experiments it is not a material source of sensitivity: replacing 10-bin ECE with the bin-free estimator mean leaves the main conclusions unchanged. Separately from these measurement axes, ordinary QA evaluation choices such as answer extraction and correctness scoring should be reported as well.

This conclusion is stable across the robustness checks: the conditioning-context effect remains larger than the estimator effect under an integer 0–100 prompt scale, non-ECE metrics, and larger Qwen2.5 Instruct checks (Appendices I, L, J, K, and N).

5 Conclusion

We show that two commonly used confidence measures, verbalized confidence and token-probability confidence, depend on measurement configuration and answer provenance. Rather than treating either scalar as a fixed property of a model-question pair, we evaluate how calibration comparisons change when the scored answer, token-probability readout, conditioning context, estimator, and elicitation provenance are varied. The sign and magnitude of the calibration gap change with the scored answer, token-probability readout, conditioning context, and elicitation provenance. In our main analysis, the conditioning-context choice shifts the gap more than the evaluation dataset does and within the same order as model family or base/Instruct variant (§4.1), while estimator choice is small.

Taken together, the results give a cautionary reading of scalar confidence claims. Verbalized confidence and token-probability scores can both be useful calibration signals, but neither should be treated as a measure of model reliability or self-knowledge independent of how it is elicited, scored, and conditioned. Comparisons between the two are therefore meaningful only when their measurement configuration is made explicit.

Limitations

Our claims concern how confidence comparisons vary with measurement configuration and answer provenance, not a universal ordering of verbalized confidence and token probability.

Model and scale coverage. The main experiments use three open 7–8B base/Instruct families. This choice makes the conditioning-context and token-probability analyses fully reproducible, but it does not cover closed models or the full range of larger chat models. Same-family Qwen2.5 checks at 14B, 32B, and 72B show the same qualitative pattern (Appendices J, K, N), but these are robustness checks rather than a systematic scale study.

Descriptive protocol comparisons. The protocol configurations differ in operational choices: which answer receives the token-probability score, how that score is read from the answer, which conditioning context is used, and which calibration estimator is applied. Pairwise differences such as $g_A - g_B$ are therefore sensitivity measurements, not estimates of how much each choice causally contributes to the observed gap. In particular, the verbalized-prefix configuration measures token probability after the prompt prefix used to elicit verbalized confidence and is best read as a conditioning-context probe rather than as a normative calibration target.

Parseability. Base models sometimes fail to follow the verbalized-confidence format, most visibly Llama-3-8B and Mistral-7B-v0.3 base, whose parse rates are low in some settings (Appendix E). In those cases, verbalized-side metrics are computed on a smaller parseable subset. Our main claims are therefore stated on the Instruct settings, where parse rates are at least 0.966; base-model results are stress tests of the measurement pipeline rather than primary evidence about verbalized calibration quality.

Prompt and provenance scope. For the main verbalized-confidence measurements, we use a continuous 0.0–1.0 probability format. An integer 0–100 check on the Instruct settings leaves the qualitative protocol-sensitivity pattern unchanged (Appendix I), but other elicitation formats and a systematic prompt-scale study remain outside our scope. The supplied-answer analysis is likewise bounded: it uses the three 7–8B Instruct models with a Qwen2.5-14B robustness check, and its wrong answers mix cross-model plausible errors with off-type fallback answers when no cross-model wrong answer is available.

Answer, decoding, and scoring conventions. The default token-probability score is assigned to the model-generated answer and read out as the answer-span geometric mean (T_1). We report gold-reference and first-token (T_2) checks separately because they change the measured quantity. All main generations use greedy decoding, and correctness is assigned by one lenient string-matching rule across datasets. Sampling-based decoding, alternative answer extraction, or stricter task-specific scoring may produce different absolute calibration values. Because all evaluations are QA-style tasks, the results may not transfer directly to long-form generation, multi-turn interaction, code generation, or decision-making settings where the answer object and correctness label are less discrete.

Ethical Considerations

This paper is a measurement study of confidence signals in existing open language models on public QA benchmarks. We do not train or deploy models, collect user data, or run experiments with human participants or crowdworkers. The experiments use TriviaQA, SciQ, TruthfulQA, and MMLU only as offline evaluation sets, and the models and datasets are used through their released artifacts and associated licenses or terms. Any released artifacts should therefore be interpreted as research artifacts for reproducing aggregate measurements; they should not be treated as a new benchmark of human knowledge or as a source of deployable confidence thresholds.

A central ethical risk is misinterpretation. Confidence scores are often used to support abstention, escalation, or other safety-relevant decisions. Our results show that verbalized confidence and token-probability confidence can change substantially with answer provenance, scored answer, token readout, and conditioning context. Reporting only a favorable scalar calibration number could encourage over-reliance on a model in high-stakes settings such as medical, legal, educational, or safety-critical QA. We therefore present the work as a protocol-sensitivity analysis and include a reporting checklist rather than recommending either signal as a standalone reliability measure. Downstream use should validate the full measurement protocol in the target domain and should not substi-

tute these confidence signals for expert review or operational risk assessment.

The public benchmarks used here may contain culturally specific coverage, named entities, outdated facts, and artifacts of their original collection processes. We do not collect new user data, and we do not release item-level personal information. Because public QA benchmarks and model-generated wrong answers may contain named entities or potentially sensitive or offensive content, we use them only for controlled offline evaluation and report aggregate results rather than reproducing individual examples. Our evaluation is English QA-style and does not assess whether calibration behavior differs across demographic groups, languages, or communities. The supplied-answer analysis also constructs plausible wrong answers from other model outputs; these outputs may include false statements and should be used only for controlled evaluation. We avoid making claims about fairness, truthfulness, or real-world safety beyond the tested setting.

The experiments require non-trivial inference compute, including larger Qwen2.5 robustness checks, although no model training or fine-tuning is performed. We mitigate avoidable environmental cost by reusing cached model outputs for most analyses, reporting implementation details needed for selective reproduction, and separating small-scale robustness checks from claims that would require extensive scaling studies.

References

Jarosław Błasiok and Preetum Nakkiran. 2024. Smooth ECE: Principled reliability diagrams via kernel smoothing. In *ICLR*. ArXiv:2309.12236.

Yuyang Dai. 2026. Rescaling confidence: What scale design reveals about LLM metacognition. ArXiv:2603.09309.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *EMNLP*, pages 295–302.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2024. A survey of confidence estimation and calibration in large language models. In *NAACL (Long Papers)*, pages 6577–6595. ArXiv:2311.08298.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *ICML*.

Chirag Gupta and Aaditya Ramdas. 2021. Distribution-free calibration guarantees for histogram binning without sample splitting. In *ICML*, pages 3942–3952. ArXiv:2105.04656.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *ICLR*. ArXiv:2009.03300.

Xinmeng Huang, Shuo Li, Mengxin Yu, Matteo Sesia, Hamed Hassani, Insup Lee, Osbert Bastani, and Edgar Dobriban. 2024. Uncertainty in language models: Assessment through rank-calibration. In *EMNLP*, pages 284–312. ArXiv:2404.03163.

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

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *ACL*. ArXiv:1705.03551.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. ArXiv:2207.05221.

Reza Khanmohammadi, Erfan Miahi, Mehrsa Mardikoraem, Simerjot Kaur, Ivan Brugere, Charese Smiley, Kundan S. Thind, and Mohammad M. Ghassemi. 2025. Calibrating LLM confidence by probing perturbed representation stability. In *EMNLP*, pages 10448–10514. ArXiv:2505.21772.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *ICML*. ArXiv:1905.00414.

Ananya Kumar, Percy S. Liang, and Tengyu Ma. 2019. Verified uncertainty calibration. In *NeurIPS (Spotlight)*. ArXiv:1909.10155.

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

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022b. TruthfulQA: Measuring how models mimic human falsehoods. In *ACL*. ArXiv:2109.07958.

Qing Lyu, Kumar Shridhar, Chaitanya Malaviya, Li Zhang, Yanai Elazar, Niket Tandon, Marianna Apidianaki, Mrinmaya Sachan, and Chris Callison-Burch. 2024. Calibrating large language models with sample consistency. ArXiv:2402.13904.

Meta AI. 2024. The llama 3 herd of models. ArXiv:2407.21783.

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

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

Qwen Team. 2024. Qwen2.5 technical report. ArXiv:2412.15115.

Rebecca Roelofs, Nicholas Cain, Jonathon Shlens, and Michael C. Mozer. 2022. Mitigating bias in calibration error estimation. In *AISTATS*. ArXiv:2012.08668.

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 *EMNLP*, pages 5433–5442. ArXiv:2305.14975.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *W-NUT*. ArXiv:1707.06209.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *ICLR*. ArXiv:2306.13063.

Daniel Yang, Yao-Hung Hubert Tsai, and Makoto Yamada. 2024. On verbalized confidence scores for LLMs. ArXiv:2412.14737.

Jize Zhang, Bhavya Kailkhura, and T. Yong-Jin Han. 2020. Mix-n-Match: Ensemble and compositional methods for uncertainty calibration in deep learning. In *ICML*. ArXiv:2003.07329.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *ICML*, pages 12697–12706. PMLR 139; arXiv:2102.09690.

A Full A–D Protocol Grid by Setting

Table 7 reports the per-setting values underlying the macro summaries in Table 2. Each row is one model-dataset setting; the main table averages these rows within the Instruct and base variant groups rather than pooling examples across settings.

B Full Three-Layer Context-Shift Measurements

Table 8 below extends §4.1: in addition to the first-token KL (L_1) discussed in the main paper, we measured two further layers per (model pair, variant, dataset, $n = 200$): a final-layer CKA-based hidden-state shift ($L_2 = 1 - \mathrm{CKA}$) and a format-aware behavioural generation-change rate (L_3). The three layers move in different directions across base and Instruct variants and correlate only weakly with one another, supporting our decision to report them side-by-side rather than as a single "drift" quantity.

C Prefix-Axis Invariance to Token Position

§3.2 raises a trivial-alternative reading of the prefix axis: that the cell-B token probability merely reflects the model's self-conditioned probability of its own committed "Guess: X" token, so the cell-A–cell-B shift would be an artifact of *which token readout* we use rather than of the prefix change. If that were the dominant explanation, the prefix-axis quantity $g_A - g_B$ would be expected to vary strongly with the token readout, including possible sign differences across T_1 and T_2. Instead, Table 9 reports broadly aligned $g_A - g_B$ values under both definitions per cell. The sign agrees on 20 of 24 cells overall and on **12 of 12 Instruct cells** (the four disagreements are all Mistral/Llama *base*, within the separately-flagged verbalized-format-failure regime). The prefix-axis sensitivity is therefore not an artifact of the T_1 answer-span readout; it is a property of the prefix change that survives the token-readout choice on every Instruct cell.

D Estimator-Axis Robustness to Bin Count

§4.1(iii) reports that the estimator axis ($g_A - g_C$, 10-bin vs. bin-free) is statistically null on Instruct. A natural concern is that this is a coincidence of the 10-bin choice. Table 10 reports raw macro point estimates after re-computing g_A with the equal-width estimator at $B \in \{10, 20, 30, 50\}$ bins (the bin-free g_C is B-independent). Across all four bin counts, g_A is stable to within 0.003 (Instruct) / 0.009 (base), and the estimator-axis quantity $g_A - g_C$ never exceeds 0.004 (Instruct) / 0.015 (base) in absolute value. The estimator axis is therefore robustly null, not a 10-bin artifact.

E Verbalized Parse Rates

The verbalized metric is defined only on samples from which a confidence is recoverable by the parser. Table 11 gives the per-cell parse rate. All 12 Instruct cells parse at ≥ 0.966, so Instruct-side metrics use essentially the full sample and the main claims (stated on Instruct) are unaffected. Base cells vary widely: Qwen-2.5-7B base parses at ≥ 0.93, but Llama-3-8B and Mistral-7B-v0.3 base parse at only 0.11–0.57, so their verbalized-side metrics are restricted to a small parseable subset while their bare-side metrics use the full sample. Base-model numbers are therefore not like-for-like and are not treated as primary calibration evidence (§5).

F Parseable-Subset Matched Analysis (Base Models)

A residual concern is that base-model gaps mix a parse-rate confound: the verbalized metric uses only parseable samples while the bare-token metric uses the full sample. Table 12 re-computes the bare-token ECE *restricted to the verbalized-parseable qids*, so both signals are on the same sample. Restricting the token side moves its ECE by only 0.022 on average, and the sign of the gap survives the matching on *all 12* base cells (0 flips). Base-model gaps are therefore not an artifact of subset mismatch; they remain excluded from our primary (Instruct) claims for the separate reason that base models frequently fail the verbalized format (§5).

G An Additional Scoring Caveat: Generated vs. Gold Token Probability

The main paper computes token probability on the model-generated answer. A further scoring choice we note here computes it on the gold reference (teacher-forced) instead, holding the correctness label fixed. Teacher forcing is widely used and not itself invalid; but *for this comparison* it changes which answer is scored: when the correctness label refers to the model-generated answer while

Model	Data	g_A	g_B	g_C	g_D
Instruct					
Llama-3-8B	MMLU	-0.034	$+0.366$	-0.035	$+0.377$
Llama-3-8B	SciQ	$+0.047$	-0.172	$+0.059$	-0.158
Llama-3-8B	TriviaQA	-0.023	-0.297	$+0.016$	-0.295
Llama-3-8B	TruthfulQA	-0.244	$+0.147$	-0.249	$+0.172$
Mistral-7B-v0.3	MMLU	$+0.054$	$+0.429$	$+0.044$	$+0.448$
Mistral-7B-v0.3	SciQ	$+0.098$	-0.076	$+0.095$	-0.073
Mistral-7B-v0.3	TriviaQA	$+0.083$	-0.321	$+0.071$	-0.328
Mistral-7B-v0.3	TruthfulQA	-0.113	$+0.443$	-0.123	$+0.446$
Qwen2.5-7B	MMLU	-0.047	$+0.379$	-0.046	$+0.391$
Qwen2.5-7B	SciQ	-0.003	-0.280	$+0.003$	-0.276
Qwen2.5-7B	TriviaQA	$+0.023$	-0.187	$+0.023$	-0.188
Qwen2.5-7B	TruthfulQA	-0.189	$+0.301$	-0.194	$+0.312$
Base					
Llama-3-8B	MMLU	$+0.049$	$+0.439$	$+0.012$	$+0.407$
Llama-3-8B	SciQ	-0.185	$+0.240$	-0.250	$+0.176$
Llama-3-8B	TriviaQA	-0.427	-0.054	-0.480	-0.096
Llama-3-8B	TruthfulQA	$+0.305$	$+0.393$	$+0.318$	$+0.378$
Mistral-7B-v0.3	MMLU	$+0.768$	$+0.813$	$+0.763$	$+0.797$
Mistral-7B-v0.3	SciQ	$+0.948$	$+0.830$	$+0.942$	$+0.819$
Mistral-7B-v0.3	TriviaQA	$+0.876$	$+0.810$	$+0.863$	$+0.794$
Mistral-7B-v0.3	TruthfulQA	$+0.683$	$+0.822$	$+0.664$	$+0.804$
Qwen2.5-7B	MMLU	-0.019	$+0.230$	-0.008	$+0.279$
Qwen2.5-7B	SciQ	-0.167	-0.297	-0.165	-0.284
Qwen2.5-7B	TriviaQA	-0.117	-0.277	-0.115	-0.265
Qwen2.5-7B	TruthfulQA	-0.036	$+0.212$	-0.040	$+0.222$

Table 7: Per-setting signed gaps for the A–D protocol grid. Negative values mean verbalized confidence has lower ECE than token probability in that setting. A is the generated-answer, bare-query, T_1, 10-bin protocol; B changes the token-probability context to the verbalized prefix; C changes the ECE estimator to the bin-free mean; D applies both changes.

the probability is read on a different, gold answer, the label and the probability describe two different scored answers, so the quantity being calibrated changes. We use this contrast as a diagnostic, not as a claim that teacher-forced gold-answer probability is invalid in all settings. It is appropriate when the target is confidence in the gold answer; it is not interchangeable with confidence in the model-generated answer. Concretely:

```
# Correct main protocol
answer = model.generate(q)
y      = score(answer, gold)
c_tok  = prob_of(answer, context=q)

# Caveat protocol
answer = model.generate(q)
# label uses generated answer
y      = score(answer, gold)

# probability uses gold answer
c_tok  = tf_prob_of(gold, context=q)
# y and c_tok now refer to
# different scored answers.
```

Table 5 in the main text summarizes the macro contrast. Figure 1 reports the per-cell sign pattern under both conventions, including the larger-model check (Qwen2.5-14B; §J).

H Per-Cell Answer-Provenance Results

Table 13 gives the per-cell verbalized confidence Y behind §4.2, for each of the twelve (Instruct model, dataset) cells. The weak correctness discrimination is consistent across cells: the self gap $Y(\text{correct}) - Y(\text{wrong})$ ranges 0.05–0.14, and the supplied-answer gap against a plausible (cross-model) distractor stays within $[-0.06, +0.10]$ on every cell. The distractor construction is also explicit: 7,229 of 11,916 supplied wrong-answer cases (60.7%) are cross-model wrong answers for the same question, and 4,687 (39.3%) are mismatched-gold off-type fallbacks. The cross-model share ranges from 39.5% to 78.5% across the twelve cells.

I Integer 0–100 Prompt-Scale Robustness Check

The main study elicits verbalized confidence with a Tian et al. (2023)-style decimal probability format in $[0, 1]$. To check that the protocol-sensitivity conclusion is not an artifact of that scale, we rerun the three Instruct model families (Llama-3-8B, Mistral-7B-v0.3, Qwen-2.5-7B; 4 datasets) with the elicitation changed to an integer 0–100 confidence and the token-probability side held fixed. This is a small robustness check on the elicitation *scale*, not a systematic prompt study (cf. §5). Table 16 reports the per-cell gap under the integer template alongside the decimal template.

Model pair	Variant	L_1 KL	L_2 (1−CKA)	L_3 behav.
Llama-3-8B	base	4.47 ± 0.88	0.41 ± 0.06	0.79 ± 0.10
	Instruct	15.29 ± 1.38	0.26 ± 0.10	0.54 ± 0.28
Mistral-7B-v0.3	base	2.60 ± 0.86	0.40 ± 0.08	1.00 ± 0.01
	Instruct	17.78 ± 0.66	0.26 ± 0.02	0.62 ± 0.22
Qwen-2.5-7B	base	3.91 ± 0.36	0.61 ± 0.07	0.75 ± 0.09
	Instruct	24.13 ± 1.68	0.36 ± 0.07	0.63 ± 0.23
Aggregate base	(n=12)	3.66 ± 1.06	0.47 ± 0.12	0.85 ± 0.13
Aggregate Instruct	(n=12)	19.07 ± 4.06	0.29 ± 0.08	0.60 ± 0.23

Table 8: Three-layer context-shift quantification, mean ± std across four datasets per (model pair, variant) cell, $n = 200$ examples per dataset. L_1: first-step next-token KL between bare and verbalized-prefixed conditioning. L_2: $1 - \mathrm{CKA}$ on final-layer hidden states. L_3: format-aware behavioral generation-change rate. The cross-cell correlations among the three layers are $r_{L_1,L_2} = -0.57$, $r_{L_1,L_3} = -0.53$, $r_{L_2,L_3} = +0.23$, so the three measurements should not be reduced to a single quantity. We treat these as descriptive diagnostics rather than causal evidence. In these measurements, L_1 is larger for Instruct variants, whereas L_2 and L_3 are lower; the pattern shows that the diagnostics capture different aspects of the context change rather than a single mechanism.

Cell	$g_A - g_B$ (T_1)	$g_A - g_B$ (T_2)
Instruct (sign agreement 12/12)		
llama3-8b/triviaqa	$+0.275$	$+0.430$
llama3-8b/sciq	$+0.218$	$+0.368$
llama3-8b/truthfulqa	-0.389	-0.247
llama3-8b/mmlu	-0.396	-0.200
mistral-7b/triviaqa	$+0.400$	$+0.438$
mistral-7b/sciq	$+0.173$	$+0.271$
mistral-7b/truthfulqa	-0.552	-0.436
mistral-7b/mmlu	-0.375	-0.295
qwen2.5-7b/triviaqa	$+0.210$	$+0.178$
qwen2.5-7b/sciq	$+0.275$	$+0.252$
qwen2.5-7b/truthfulqa	-0.489	-0.503
qwen2.5-7b/mmlu	-0.424	-0.449
Base (sign agreement 8/12; verbalized-format-failure regime)		
macro -0.125 (T_1) vs. -0.068 (T_2)		

Table 9: Prefix-axis shift $g_A - g_B$ under two token-probability readouts (T_1 = answer-span geometric mean; T_2 = first answer-token probability). Sign agreement is 12/12 on Instruct cells, showing that the prefix-axis effect is not an artifact of the T_1 answer-span readout alone. Under the self-conditioning reading of §3.2, stronger readout dependence would be expected. Bootstrap means, $B = 200$.

J Larger Open Instruct Robustness Check (Qwen2.5-14B)

To check whether the protocol sensitivity is specific to the 7–8B regime, we run one larger open Instruct model, `Qwen2.5-14B-Instruct`, with the same generated-answer protocol (four datasets, decimal verbalized-confidence format; the integer template of §I is also collected). A single model cannot establish scale generality, and we make no scaling claim; this check only tests whether the 7–8B findings are a small-scale artifact. The pattern is the same as on the 7–8B pairs. The prefix axis is again the dominant, sign-destabilizing axis (mean absolute shift ≈ 0.37 across the four datasets, flipping the sign of the gap on $2/4$ cells: MMLU and TruthfulQA move from verbalized-better to token-better), while the estimator axis is again null (mean shift ≈ -0.0005, ≤ 0.004 on every cell). The token-readout axis is the weaker axis here ($g(T_1) - g(T_2) \approx +0.04$ on average, comparable in magnitude to the 7–8B Instruct cells but with no cell-level sign disagreement). The large ECE reduction again does not replicate: $0/4$ cells reach $\rho \leq 0.55$ under the decimal template ($\rho \approx 0.66$–0.92; verbalized modestly better, far from a $2\times$ reduction). Parse rates are high (≥ 0.999 on all four datasets), so no format-failure confound applies. The generated-vs-gold scored-answer check also shows the same scored-answer sensitivity (Table 20).

B	Instruct (n=12) g_A	Instruct (n=12) $g_A - g_C$	base (n=12) g_A	base (n=12) $g_A - g_C$
10	-0.029	-0.001	$+0.223$	$+0.015$
20	-0.030	-0.002	$+0.219$	$+0.010$
30	-0.031	-0.003	$+0.217$	$+0.008$
50	-0.032	-0.004	$+0.214$	$+0.005$

Table 10: ECE bin-sweep. g_A is the bare-prefix gap computed with the equal-width estimator at B bins; g_C is the bin-free estimator-mean gap (B-independent, -0.028 Instruct / $+0.209$ base). g_A varies by ≤ 0.009 across all bin counts and the estimator axis $g_A - g_C$ stays within ± 0.015, confirming the §4.1(iii) null is not specific to $B = 10$. Raw macro point estimates across 12 cells per variant; these can differ slightly from the bootstrap means in Table 2.

K Larger Open Instruct Robustness Check (Qwen2.5-32B)

We additionally run `Qwen2.5-32B-Instruct` under the same generated-answer protocol (four datasets, decimal verbalized-confidence format), filling the gap between the 14B and 72B checks. As with the 14B model this is a single-model robustness check, not a scaling study, and we make no scale-generality claim. The pattern is again the same as on the 7–8B pairs. The prefix axis is the dominant, sign-destabilizing axis (mean absolute shift ≈ 0.38 across the four datasets, flipping the sign of the gap on $2/4$ cells: MMLU and TruthfulQA move from verbalized-better to token-better), while the estimator axis is null (mean shift $\approx +0.002$, ≤ 0.003 on every cell). The token-readout axis is again the weaker axis ($g(T_1) - g(T_2) \approx +0.03$ on average, with no cell-level sign disagreement). The large ECE reduction does not replicate: $0/4$ cells reach $\rho \leq 0.55$ ($\rho \approx 0.71$–0.93). Parse rates are 1.000 on all four datasets, so no format-failure confound applies. The generated-vs-gold scored-answer check also shows the same scored-answer sensitivity (Table 20).

L Multi-Metric Robustness

To test whether the protocol sensitivity is an artifact of 10-bin ECE, we recompute the verbalized-vs-token gap under four additional scoring rules on the same 24 cells, purely from the stored per-example predictions (no model reload): the Brier score, a clipped negative log-likelihood (ϵ=10^{-6}), AUROC for correctness prediction, and the area under the risk–coverage curve (AURC), a selective-risk/abstention utility. For metric M we form

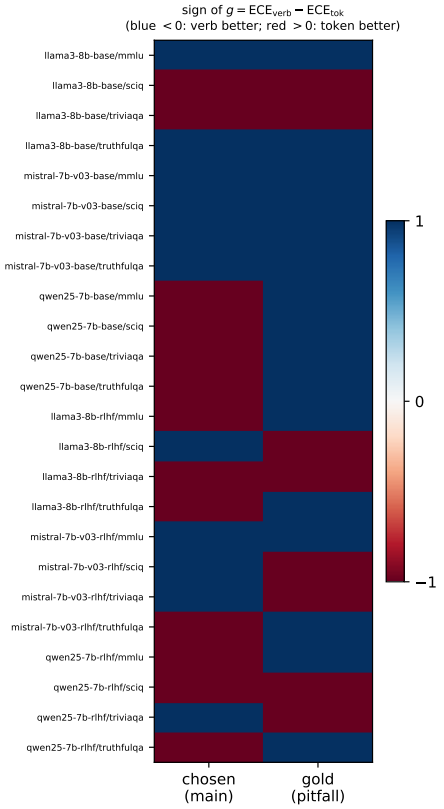

Figure 1: Scoring token probability on the gold answer instead of the model-generated answer flips the conclusion on most Instruct cells. Sign of the verbalized-vs-token gap g per (model, variant, dataset) cell under generated-answer (main) vs. gold-teacher-forced token-probability scoring, correctness held fixed. Blue ($g < 0$): verbalized better; red ($g > 0$): token-probability better. A column-wise color change indicates the scoring choice flips the conclusion (13/24 cells; 9/12 Instruct).

the same per-cell gap as in the main grid, $g_A^M = M(c_{\text{verb}}) - M(c_{\text{tok}}; \text{bare})$ and g_B^M under the verbalized prefix, and report the prefix shift $g_A^M - g_B^M$ (Table 19). The estimator axis (10-bin vs. bin-free) exists only for ECE; proper scores and ranking metrics have no binning, so it is not applicable for them by construction. The *scored-answer critique* behind the generated-vs-gold caveat is metric-agnostic: it changes which answer the token probability is assigned to before entering any metric. The *exact sign-flip counts*, however, are empirical and may differ under Brier, NLL, AUROC, or AURC. We therefore use the ECE generated-vs-gold experiment ($13/24$ cells; Appendix G) as a concrete demonstration of the scored-answer caveat, while the multi-metric analysis here tests the separate question of whether the prefix sensitivity is specific to 10-bin ECE; we do not recompute the gold-

Model pair	Variant	TriviaQA	SciQ	TruthfulQA	MMLU
Llama-3-8B	base	0.24	0.12	0.11	0.16
	Instruct	1.00	1.00	1.00	1.00
Mistral-7B-v0.3	base	0.22	0.57	0.32	0.36
	Instruct	1.00	0.99	0.97	0.97
Qwen-2.5-7B	base	0.98	0.99	1.00	0.93
	Instruct	1.00	1.00	1.00	1.00

Table 11: Per-cell verbalized parse rate (fraction of samples from which a confidence value is recovered). Instruct cells: min 0.966 (Mistral-7B-v0.3-Instruct / TruthfulQA), all ≥ 0.96. Llama-3-8B and Mistral-7B-v0.3 *base* parse at 0.11–0.57; Qwen-2.5-7B base parses well (≥ 0.93).

Model	Data	parse	tok_{full}	tok_{sub}	g_{full}	g_{sub}
Llama-3-8B	TriviaQA	0.24	0.648	0.731	-0.43	-0.51
	SciQ	0.12	0.586	0.636	-0.19	-0.24
	TruthfulQA	0.11	0.168	0.229	$+0.31$	$+0.24$
	MMLU	0.16	0.547	0.504	$+0.05$	$+0.09$
Mistral-7B	TriviaQA	0.22	0.056	0.056	$+0.88$	$+0.88$
	SciQ	0.57	0.022	0.029	$+0.95$	$+0.94$
	TruthfulQA	0.32	0.275	0.290	$+0.68$	$+0.67$
	MMLU	0.36	0.202	0.204	$+0.77$	$+0.77$
Qwen2.5-7B	TriviaQA	0.98	0.372	0.372	-0.12	-0.12
	SciQ	0.99	0.419	0.417	-0.17	-0.17
	TruthfulQA	1.00	0.517	0.517	-0.04	-0.04
	MMLU	0.93	0.528	0.524	-0.02	-0.02

Table 12: Base-model bare-token ECE on the full sample (tok_{full}) vs. restricted to the verbalized-parseable subset (tok_{sub}), and the resulting gaps. Mean $|\text{tok}_{\text{sub}} - \text{tok}_{\text{full}}| = 0.022$; the gap sign is unchanged on all 12 cells.

reference token probability under non-ECE metrics, as that requires an additional token-probability extraction pass. Huang et al. (2024) make a related point that confidence signals on different scales should be compared with rank-based diagnostics.

M Dataset-Specific Correctness Scoring

All four datasets are scored by the single lenient normalized string-match rule of §3 (lowercase; non-alphanumeric → space; whitespace collapsed; correct iff normalized strings are equal or one contains the other). Table 21 gives the split, example count, answer format, gold object, and parse-failure handling per dataset. SciQ and MMLU are multiple-choice sources but are scored as *free-form short text against the correct option string*, not as letter options, so the same rule and the same answer-span extraction apply throughout. The MMLU subset is four subjects with per-subject counts capped by dataset availability (high-school physics 151, philosophy 311, high-school government and politics 193, miscellaneous 500; 1,155 total), not a uniform 500 per subject. Verbalized rows with no parseable confidence are excluded from that cell's verbalized-side ECE; each side is then computed on its own valid sample (the bare/token side on the full sample, the verbalized side on its parseable subset, not intersected by `qid`), which is immaterial for the high-parse-rate Instruct cells and is checked for base models by the `qid`-matched analysis of Appendix F. Per-cell parse rates are in Appendix E.

Scorer-sensitivity check. The rule above is lenient: it credits a normalized exact match *or* two-way substring containment. Correctness scoring is itself a protocol choice, so we re-derive every `correct` label under the stricter *normalized-exact-match-only* rule (dropping the substring branch) and recompute the verbalized-vs-token gaps from the stored predictions (Table 22; no model reload). The stricter scorer changes absolute magnitudes substantially, but the qualitative ordering is unchanged: the conditioning-context shift remains much larger than the estimator shift. The headline numbers in the main text are reported under the lenient rule; a systematic study of scorer design is left to future work.

N Large-Model Specificity Diagnostic (Qwen2.5-72B)

The 7–8B pairs and the 14B/32B checks (§J, §K) leave open whether the protocol sensitivity is specific to sub-32B models. As a specificity check—*not* a scaling study—we run one larger open

Model	Data	self		supplied			discrim.	
		Y_{cor}	Y_{wr}	gold	plaus.	off-type	self	suppl.
Llama-3-8B	TriviaQA	0.870	0.796	0.812	0.772	0.377	+0.073	+0.015
	SciQ	0.876	0.825	0.855	0.827	0.320	+0.051	+0.004
	TruthfulQA	0.775	0.655	0.736	0.715	0.377	+0.120	+0.010
	MMLU	0.843	0.778	0.797	0.788	0.409	+0.065	−0.008
Mistral-7B-v0.3	TriviaQA	0.957	0.838	0.796	0.732	0.124	+0.119	+0.018
	SciQ	0.970	0.894	0.938	0.892	0.121	+0.076	+0.020
	TruthfulQA	0.922	0.780	0.793	0.689	0.401	+0.142	+0.093
	MMLU	0.924	0.868	0.870	0.858	0.204	+0.056	−0.000
Qwen-2.5-7B	TriviaQA	0.938	0.810	0.802	0.818	0.197	+0.128	−0.059
	SciQ	0.942	0.861	0.908	0.815	0.115	+0.081	+0.060
	TruthfulQA	0.850	0.766	0.809	0.706	0.352	+0.084	+0.095
	MMLU	0.925	0.832	0.834	0.800	0.212	+0.093	+0.007

Table 13: Per-cell mean verbalized confidence Y by `Guess:`-slot fill, twelve Instruct (model, dataset) cells. "self": the model's own answer, split by correctness; "supplied": a supplied gold, plausible cross-model wrong, or off-type wrong answer. "self" discrimination is $Y_{\text{cor}} - Y_{\text{wr}}$; "suppl." is the paired $Y_{\text{gold}} - Y_{\text{plaus.}}$. Across all twelve cells the supplied-answer discrimination stays near zero while off-type answers are rejected.

Paired contrast	mean [95% CI]
gold − plaus. wrong	+0.021 [+0.014, +0.028]
self − gold, diff. strings	+0.055 [+0.046, +0.064]
self − gold, exact match	+0.000 [−0.001, +0.001]
gold − self, initially wrong	−0.010 [−0.016, −0.004]

Table 14: Bootstrap intervals for the paired contrasts in Table 6. All rows use the main 7–8B Instruct supplied-answer analysis.

Data	gold	plaus.	off-type	diff. pref.
TriviaQA	0.719	0.634	0.084	+0.193
SciQ	0.895	0.747	0.069	+0.019
TruthfulQA	0.763	0.683	0.050	−0.002
MMLU	0.820	0.756	0.080	−0.009
Macro	0.799	0.705	0.071	–
Pooled	–	–	–	+0.104 [+0.087, +0.122]

Table 15: `Qwen2.5-14B-Instruct` supplied-answer robustness results. The first three numeric columns are mean verbalized confidences under supplied `Guess:` conditions. "diff. pref." is the paired self-minus-gold contrast on correct answers whose strings differ; the main text reports the pooled contrast.

Axis / metric (Instruct, n=12)	decimal	integer 0–100
macro g_A (generated, bare)	−0.029	−0.033
prefix shift ($g_A - g_B$, macro)	−0.090	−0.112
estimator shift ($g_A - g_C$, macro)	−0.001	−0.002
Tian $\rho \leq 0.55$ cells	0/12	0/12

Table 16: Changing the verbalized elicitation from a decimal probability to an integer 0–100 leaves the protocol-sensitivity pattern qualitatively unchanged: the prefix shift stays an order of magnitude larger than the estimator shift (mean absolute per-cell prefix shift $0.35 \rightarrow 0.27$; $6/12$ cells sign-flip under the integer template), the estimator axis stays null, and the $2\times$ magnitude does not replicate on the Instruct cells under either template.

Instruct model, `Qwen2.5-72B-Instruct`, loaded full-precision in BF16 (the model's native dtype) and sharded across four 80GB A100 GPUs via `device_map="auto"`; token logprobs are computed exactly. We evaluate the same four datasets as the 14B/32B checks (MMLU at full 1,155 examples; TriviaQA, SciQ, and TruthfulQA at 250-example subsets retained from the original diagnostic design), using the generated-answer protocol, the decimal verbalized-confidence format, and the 10-bin estimator. A single larger model cannot establish scale generality and we make no scaling claim; this only tests whether the 7–32B pattern disappears at 72B. It does not (Table 23). The prefix axis is again the dominant, sign-destabilizing axis (mean absolute shift ≈ 0.41 across the four datasets, flipping the sign on MMLU and TruthfulQA, exactly as in the 14B and 32B checks); the estimator axis is again null ($g_A - g_C$ macro ≈ 0.000, ≤ 0.008 per cell); and the large ECE reduction again does not replicate ($0/4$ cells reach $\rho \leq 0.55$; $\rho \approx 0.63$–0.88). Parse rates are ≥ 0.998 on all four datasets. The generated-vs-gold scored-answer check again shows the same scored-answer sensitivity (Table 20), with sign flips on MMLU and TruthfulQA.

Dataset	g_A	g_B	g_C	ρ
MMLU	-0.050	$+0.281$	-0.051	0.92
SciQ	-0.049	-0.391	-0.047	0.84
TriviaQA	-0.092	-0.482	-0.092	0.66
TruthfulQA	-0.200	$+0.215$	-0.204	0.71

Table 17: `Qwen2.5-14B-Instruct`: per-dataset verbalized-vs-token gap under the generated-answer bare cell (g_A), the verbalized-prefix cell (g_B), the bin-free estimator (g_C), and the Tian magnitude ratio $\rho = \mathrm{ECE}(c_{\text{verb}})/\mathrm{ECE}(c_{\text{tok}}; \text{bare})$. Prefix moves the gap by an order of magnitude more than the estimator and flips its sign on MMLU and TruthfulQA; $\rho > 0.55$ on all four cells. Same qualitative pattern as the 7–8B pairs; reported as a single-model robustness check, not a scaling study.

Dataset	g_A	g_B	g_C	ρ
MMLU	-0.042	$+0.282$	-0.044	0.93
SciQ	-0.051	-0.415	-0.053	0.83
TriviaQA	-0.083	-0.483	-0.084	0.71
TruthfulQA	-0.205	$+0.233$	-0.207	0.72

Table 18: `Qwen2.5-32B-Instruct`: per-dataset verbalized-vs-token gap under the generated-answer bare cell (g_A), the verbalized-prefix cell (g_B), the bin-free estimator (g_C), and the Tian magnitude ratio $\rho = \mathrm{ECE}(c_{\text{verb}})/\mathrm{ECE}(c_{\text{tok}}; \text{bare})$. Prefix moves the gap by an order of magnitude more than the estimator and flips its sign on MMLU and TruthfulQA; $\rho > 0.55$ on all four cells. Same qualitative pattern as the 7–8B and 14B models; reported as a single-model robustness check, not a scaling study.

Metric (Instruct, n=12)	macro	mean$\lvert\Delta\rvert$	sign-flip
ECE (10-bin)	-0.090	0.350	9/12
Brier	-0.047	0.262	9/12
NLL (clipped)	$+1.141$	1.192	5/12
AUROC	-0.196	0.197	1/12
AURC	$+0.043$	0.084	2/12

Table 19: Prefix shift $g_A^M - g_B^M$ on the twelve Instruct cells under five scoring rules. The prefix choice is a first-order source of variation under *every* metric—not only 10-bin ECE—with a large mean absolute per-cell shift; the absolute ranking of which signal looks better does vary by metric (e.g. NLL is dominated by near-0/1 probabilities), which is expected and is precisely why the comparison must specify its protocol. The estimator axis is ECE-only and null (Appendix D); proper/ranking scores have no binning.

Model	g_{gen}	g_{gold}	sign-flip
Qwen2.5-14B	-0.098	$+0.000$	2/4
Qwen2.5-32B	-0.095	$+0.018$	2/4
Qwen2.5-72B	-0.129	-0.004	2/4

Table 20: Generated-answer vs. gold-teacher-forced token-probability scoring for larger Qwen2.5-Instruct robustness checks. g_{gen} is the main generated-answer gap and g_{gold} scores the gold reference while holding the generated-answer correctness label fixed.

Dataset	Split / n	Format & gold
TriviaQA	`rc.nocontext` val, 1,000	free-form; answer aliases
SciQ	`test`, 1,000	MC source, scored as short text; correct option
TruthfulQA	`generation` val, 817	free-form; reference correct set
MMLU	`test`, 1,155	MC source, scored as short text; correct option

Table 21: Dataset-specific scoring. One normalized string rule for all; SciQ/MMLU scored as free-form short text (not letter options). Parse-failure rows are excluded from the verbalized-side ECE (not scored as wrong); each side uses its own valid sample (§3).

Scorer	g_A	$g_A - g_B$	$g_A - g_C$	mean $\lvert\Delta g_A\rvert$
Lenient (main)	-0.029	-0.090	-0.001	–
Exact match only	-0.205	-0.500	-0.000	0.26

Table 22: Scorer-sensitivity check on the twelve Instruct settings. The strict scorer uses normalized exact match only. The final column is the mean per-setting absolute change in g_A relative to the lenient main scorer.

Dataset	n	g_A	g_B	g_C	ρ
MMLU	1,155	-0.081	$+0.276$	-0.082	0.88
TriviaQA	250	-0.055	-0.422	-0.057	0.82
SciQ	250	-0.075	-0.509	-0.067	0.74
TruthfulQA	250	-0.305	$+0.189$	-0.310	0.63

Table 23: `Qwen2.5-72B-Instruct` (BF16 full precision): verbalized-vs-token gap under the generated-answer bare cell (g_A), the verbalized-prefix cell (g_B), the bin-free estimator (g_C), and the Tian magnitude ratio ρ. MMLU is at the full 1,155 examples; the other three datasets retain 250-example subsets from the original diagnostic design. Prefix moves the gap by an order of magnitude more than the estimator and flips its sign on MMLU and TruthfulQA; $\rho > 0.55$ on all four cells. Same qualitative pattern as the 7–8B, 14B, and 32B models; reported as a single-model specificity check, not a scaling study.