RELIABILITY AND INFORMATIVENESS OF VERBALIZED CONFIDENCE IN INSTRUCTION-TUNED LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Language models are increasingly asked to state how confident they are, and the numbers they give back are reported as calibration evidence. Calibration presupposes a measurement, however: a quantity that carries information about correctness and that stays put under changes to how it is requested. We test both properties across open-weight instruction-tuned models spanning more than an order of magnitude in scale, on five English benchmarks, under five prompt formulations and seven ways of asking for a confidence value. Verbalized confidence separates correct answers from incorrect ones only slightly better than chance, with a mean AUROC of 0.536 against 0.682 for the first-token probability of the same responses. It is also unstable. Holding the model, the question and the model's own answer fixed and changing only the wording of the request, the two formulations used in prior work barely agree, and a model asked how likely its answer is to be wrong does not give a value consistent with what it reports when asked how likely that answer is to be right. Often no value is recorded at all, and much of that missingness belongs to the extraction rule rather than to the model. None of this resolves with scale: within one family at fixed training recipe, accuracy and first-token informativeness both rise steeply across a sixty-six-fold parameter range while the verbalized report stays near chance. A model trained to verbalize confidence behaves differently under the same procedure, which indicates that our measurement registers a trained signal where one is present. The practical consequences follow. Which model looks best calibrated changes with the wording of the request in roughly a third of our comparisons, and abstention built on this signal recovers about a seventh of the accuracy available from token probability. We recommend that checks of informativeness and reliability accompany any reported verbal calibration number.

1 INTRODUCTION

Deployed language models are routinely asked not only for an answer but for some indication of how much that answer should be trusted. Two signals are available. The first is the probability the model assigns to its own output tokens. The second is a number it states in natural language when asked how confident it is. The second has become popular for reasons with little to do with its measurement properties: it needs no access to logits, it survives an API boundary, it applies to free-form output where token probabilities are awkward to define, and it is legible to a human reader.

A substantial literature evaluates this signal by asking whether it is calibrated, typically through expected calibration error against observed accuracy. Reported calibration is usually poor, and several recent methods train models to verbalize better-calibrated confidence (Li et al., 2025; Damani et al., 2026; Ma et al., 2026). The premise common to that work is that verbalized confidence is a quantity with a calibration property that training can improve.

A prior question has been skipped. Calibration is a property of a measurement, and before asking whether a measurement is accurate one asks whether it is a measurement at all: whether it carries information about the thing it is supposed to track, and whether repeating it under conditions that should not matter returns the same value. These are the standard notions of validity and reliability,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

and they are ordered. An instrument that disagrees with itself cannot be calibrated in any useful sense, because its calibration error is then a property of the particular administration rather than of the model. This paper applies those checks to verbalized confidence in instruction-tuned models, and then asks whether the answers change with scale.

Three properties fail together. Verbalized confidence barely separates correct answers from incorrect ones; it does not agree with itself when the request is reworded, nor with its own answer to the complementary question; and it is often not obtainable at all, with parse rates spanning the full range from 0.00 to 1.00 depending on the model, the prompt and the phrasing. Neither scale nor instruction tuning repairs this. Within one family at fixed training recipe, accuracy and first-token informativeness both rise steeply across a sixty-six-fold range in parameters while the verbalized report does not, and instruction tuning raises the rate at which a number can be parsed, in one case from 0.032 to 0.993, without moving AUROC away from chance. Two consequences bear on published practice: the ordering of models by verbal calibration error changes with the elicitation wording in roughly half the comparisons we ran, and abstention built on this signal recovers about a seventh of the accuracy available from the token probability obtainable from the same model with logit access.

Our contributions are as follows.

1. We measure informativeness, test-retest reliability and coherence of verbalized confidence across 14 instruction-tuned models, five benchmarks, five prompt formulations and seven elicitation phrasings, and find all three close to the floor. To our knowledge neither reliability under paraphrase of the request nor coherence against the complementary question has previously been reported for this signal.
2. We show that measurability is protocol-dependent, varying from 0.00 to 1.00 with both the elicitation wording and the extraction pattern, so a reported verbal calibration number is conditional on a subset the protocol helped select.
3. We give a controlled scaling result over a sixty-six-fold parameter range at fixed family and recipe, and quantify the downstream cost in ranking instability and in selective prediction. Per-setting outputs and the analysis code are released.

We are not claiming that verbalized confidence is unusable in general. Methods that train the signal explicitly are the obvious counterexample, and we use one as a positive control: a ConfTuner checkpoint and its untuned base run through the same pipeline give verbalized AUROC of 0.605 against 0.517, so our measurement detects a trained verbal signal where one is present. Our claims concern instruction-tuned models not trained for this purpose, in the size range we study, on English multiple-choice and classification tasks.

2 Related Work

Eliciting confidence in words. Lin et al. (2022) showed that models can be trained to express uncertainty in language, and Kadavath et al. (2022) that they carry some internal signal about whether their answers are correct. Tian et al. (2023) and Xiong et al. (2024) established the prompting recipes now in common use, asking for a probability or a percentage after the answer, and both reported systematic overconfidence. We adopt their two phrasings unchanged, and the pair forms the basis of our reliability comparison.

Training the signal. Several recent methods optimise elicited confidence directly: ConfTuner (Li et al., 2025) fine-tunes with a tokenized Brier score, RLCR (Damani et al., 2026) adds a calibration term to a reinforcement learning reward, and DCPO (Ma et al., 2026) decouples the reasoning and confidence objectives. This line of work assumes what we test, and our results are consistent with its premise rather than opposed to it: we use a released ConfTuner checkpoint as a positive control, and the improvement it shows is what lets us attribute the null result elsewhere to the models rather than to our instrument.

Calibration metrics and measurement. Expected calibration error (Guo et al., 2017; Naeini et al., 2015) compares average confidence to average accuracy within bins, and is well defined whether or not the input orders examples usefully, so a nearly constant signal still produces a finite ECE. Our use of AUROC alongside it follows from that. Jacobs & Wallach (2021) argue that a metric encodes

an unstated theory of what it measures and that the gap is testable, and Raji et al. (2021) make a related argument about benchmarks; we run those tests on one widely reported quantity.

Protocol sensitivity. Reported results move under changes that should not matter: models do not use instructions as literal instruction-following would predict (Webson & Pavlick, 2022), formatting alone swings accuracy (Sclar et al., 2024), multi-prompt evaluation has been argued for on those grounds (Mizrahi et al., 2024), leaderboard orderings shift under small perturbations (Alzahrani et al., 2024), and harness details change published numbers (Biderman et al., 2026). That work establishes that accuracy is protocol-sensitive. We extend it by asking whether one reported quantity is a measurement at all.

3 Experimental Setup

3.1 Models, benchmarks and prompts

We evaluate 21 open-weight checkpoints spanning eight model families and roughly 0.5B to 33B parameters: 16 instruction-tuned models and five base models paired with their instruction-tuned counterparts. Two of the instruction-tuned checkpoints serve only as the positive control of Section 3.3. Three groupings are used. Our *main set* of ten instruction-tuned models between 1.0B and 7.6B, drawn from the Llama (Meta AI, 2024), Gemma (Gemma Team, 2025), SmolLM (Laborie & Callan-Jones, 2025), Qwen (Qwen et al., 2025), Phi (Microsoft et al., 2025), Mistral and DeepSeek (DeepSeek-AI, 2025) families, carries the main informativeness analysis. A *five-model subset* is used wherever an analysis requires repeating the elicitation, which is more expensive. A *scaling series* of six Qwen-2.5-Instruct models from 0.5B to 32.8B holds family, tokenizer and training recipe fixed so that parameter count varies alone. The five base and instruct pairs isolate instruction tuning. Exact checkpoint identifiers, sizes and group memberships are listed in Appendix A.

Five English benchmarks are used, chosen so that correctness is unambiguous and the label set closed: SST-2 (Socher et al., 2013), MNLI (Williams et al., 2018), AG News (Zhang et al., 2015), ARC-Challenge (Clark et al., 2018) and MMLU-Pro (Wang et al., 2024), covering 2-, 3-, 4- and 10-way decisions. Every combination of a model, a benchmark and a prompt formulation is one *evaluation setting*, the unit of analysis throughout. Each setting uses the same 500 items, drawn once per benchmark with a fixed seed and reused across every model, prompt formulation and way of requesting confidence, so all comparisons are paired at the item level.

Five prompt formulations are applied to every item, each perturbing one aspect of the request while leaving the task and the gold label untouched: surface paraphrasing, instruction reordering, three-shot prompting, a change of output-format instruction, and implicit framing, in which the task instruction is omitted altogether. Templates are reproduced in Appendix B. Analyses of verbalized confidence use the three formulations that leave both the task regime and the output format intact, namely paraphrasing, reordering and implicit framing. Three-shot prompting changes the task regime, and the output-format change induces chain-of-thought on reasoning tasks, which interacts with answer extraction.

3.2 Confidence signals and measures

First-token probability. For multiple-choice items we issue a dedicated single-token generation and read the probability assigned to the predicted choice letter, renormalized over the valid choice letters in the returned top-k window ($k = 200$). A dedicated call avoids the ambiguity that arises when the first step of a free-text generation is not a label token.

Verbalized confidence. After the model answers, a suffix requesting a confidence value is appended to the same context and eight tokens are generated. The parser tries three patterns in order: a decimal in $[0, 1]$ anywhere in the text, then an integer followed by a percent sign or the word *percent*, then a bare integer in $[0, 100]$, accepted only when it is the entire generation. The first two take the first match; the third does not fire on an integer embedded in a sentence, a property that turns out to be part of the measurement (Section 4.2). Verbal parse rate (VPR) is the fraction of items in a setting that yield a value, and we require at least 0.80 for a setting to enter any mean over verbalized confidence. Seven elicitation suffixes are used: the decimal request of Tian et al. (2023) and the percentage request of Xiong et al. (2024), which parse at essentially the same rate and carry the

primary reliability analysis, a betting framing and a terse fill-in form, and three added later with an explicit format instruction and a worked example, one of which asks for the probability that the answer is *wrong* and supports the coherence test. Full text is in Appendix B.

Measures. *Informativeness* is the area under the ROC curve for the confidence signal predicting item-level correctness within a setting; AUROC is threshold-free and invariant to monotone rescaling, so it separates a signal that is miscalibrated but ordered from one carrying no ordering information, and ties are resolved by mid-rank. *Reliability* is the agreement between two phrasings on the same item with the model's answer held fixed: only the elicitation step is recomputed, the stored answer being spliced back into the context before the suffix is appended. We report correlation on each pair's shared-parse subset together with the standard deviation of the phrasing-induced difference relative to that of the signal. *Coherence* does not depend on judging two phrasings equivalent: the inverted phrasing elicits $P(\text{wrong})$, so a model treating its stated values as probabilities should satisfy $P(\text{wrong}) + P(\text{correct}) = 1$ on the same item, and we report the correlation between $1 - P(\text{wrong})$ and the stated $P(\text{correct})$, the mean of their sum, and the share of items whose sum departs from one. *Calibration* is expected calibration error with ten equal-width bins, reported for continuity with prior work rather than as our primary measure. *Variance attribution* decomposes each metric's variation within a benchmark across model identity and prompt formulation using sequential η^2; running it within each benchmark keeps model-by-benchmark interaction from being charged to the protocol. Appendix H gives both in full.

3.3 Positive control

A null result about a measurement invites the objection that the measurement is at fault. We therefore include a checkpoint trained to verbalize calibrated confidence: ConfTuner (Li et al., 2025), which fine-tunes with a tokenized Brier score loss, in its released Ministral-8B form, together with the untuned Ministral-8B-Instruct-2410 it was trained from. The pair differs only in the confidence tuning and both were run under the settings used for every other model. Two caveats: our benchmarks are not the ones the checkpoint was evaluated on, so the comparison is out of distribution for it and understates the method; and the checkpoint is Mistral-family, so it controls for the pipeline rather than providing a matched comparison against the other families we study.

3.4 Protocol and reproducibility

All inference uses vLLM (Kwon et al., 2023) with greedy decoding, in bfloat16 throughout the main corpus. The repeated elicitation of Section 4.3 is the one exception: the original four phrasings were replayed in float16 on a device without bfloat16 support and the three later ones in bfloat16, so those blocks are analysed separately and never pooled. Confidence intervals are 95% percentile bootstrap with 1000 resamples and a fixed seed, the resampling unit matched to the estimand; Appendix H gives the procedure in full.

Every accuracy in this paper uses one rule, which we call predicted-letter extraction: the first character of the first non-empty line if it is a valid choice letter, otherwise a reverse scan of the lines for a standalone letter. Appendix G names it in the released code and compares it against seven alternatives. An earlier implementation of this work scored accuracy differently and without documenting it, marking an item correct when the generation, or its last non-empty line, begins with the gold letter. The two agree closely on the three formulations analysed here, within 0.01 of each other, and diverge on the chain-of-thought formulation, where predicted-letter extraction gives 0.552 against 0.449. We rescored the entire corpus under the one rule. Parse rate does not depend on this choice, but AUROC does, since the correctness label it scores against is produced by it, so Appendix F recomputes both signals under every complete rule: verbalized AUROC stays between 0.527 and 0.535 and token AUROC between 0.641 and 0.676, so the comparison is not an artefact of the rule. The complete model outputs behind every scored item, the repeated elicitations, and the analysis code are released at [anonymized].

4 Results

Unless stated otherwise, results cover the ten-model main set on the three prompt formulations that leave task regime and output format intact, and use only settings clearing the parse-rate threshold of

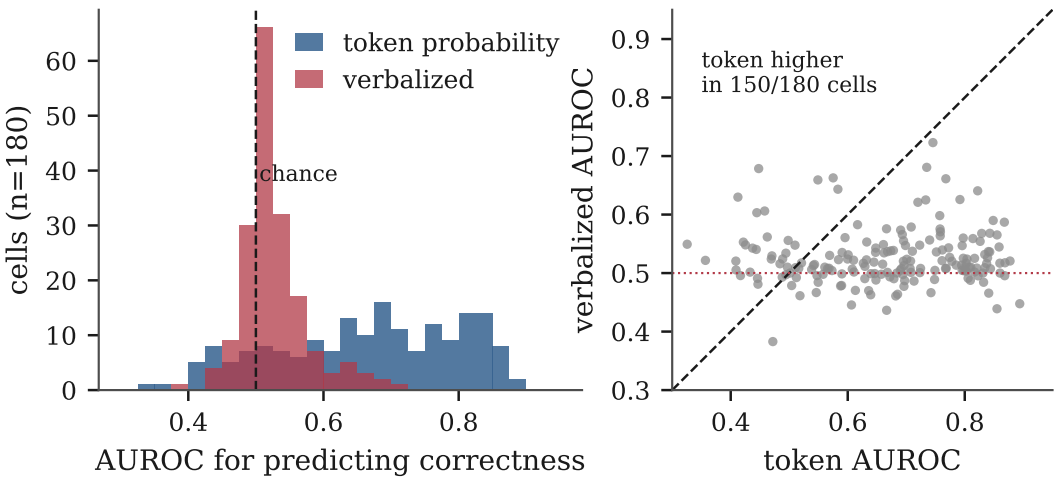

Figure 1: Informativeness over the 102 evaluation settings clearing the parse threshold. Left: distribution of AUROC for predicting item-level correctness; the verbalized signal concentrates at chance while first-token probability extends past 0.85. Right: the same settings paired, the diagonal marking equal performance and the dotted line chance for the verbalized signal.

0.80. Counts differ between analyses because the threshold excludes different settings for different signals, and we give the count with each result.

4.1 Verbalized confidence carries little information about correctness

Verbalized confidence carries little information about whether an answer is correct. Across the 102 evaluation settings that meet our parse-rate criterion, its mean AUROC is 0.536, barely above chance, against 0.682 for the first-token probability of the same responses. Token probability is the better ordering in 89 of those settings (Figure 1). The same pattern holds when all five prompt formulations are admitted, and in every model taken on its own; because settings share items and models, we also bootstrap by resampling whole models, which leaves the gap comfortably clear of zero (Table 2; Appendix F).

The reason is visible in the distribution of the values themselves. One value accounts for 76.0% of responses in a typical setting, and the median setting uses five distinct values across 500 items. Pooling the main set, three values (0.9, 0.95 and 0.8) account for 76.2% of 57,089 parsed responses. The signal is close to constant, so there is little ordering information available for any downstream use to exploit.

This has a direct consequence for how the quantity is usually summarised. Because expected calibration error compares mean confidence to mean accuracy within bins, a near-constant confidence makes ECE approximately a function of accuracy alone. Across the same settings we find a correlation of 0.923 between verbal ECE and one minus accuracy, with a mean absolute difference of 0.111 between the two. The corresponding correlation for token ECE is 0.439. Reporting verbal ECE for these models therefore adds little beyond the accuracy already reported alongside it.

4.2 Whether the quantity can be measured depends on the protocol

Whether a confidence value can be recovered at all is itself highly protocol dependent. Some model and prompt combinations yield a usable value almost every time and others almost never, the full range from 0.000 to 1.000 being attested. Nearly a third of the settings in the main set fall below our parse-rate threshold, and the failures concentrate in particular combinations rather than in a few bad prompts: prompt formulation has a small main effect on parse rate while the model-by-prompt interaction is roughly ten times larger (Appendix E).

Elicitation wording matters at least as much. Holding model and items fixed and varying only the suffix, mean parse rate across the five primary models ranges from 0.155 for a betting framing to 0.915 for the decimal request of Tian et al. (2023), with the percentage request of Xiong et al. (2024) at 0.894 and a terse fill-in form at 0.523. Two phrasings drawn from prior work are almost always parseable; two a practitioner might reasonably write are frequently not. Any reported verbal calibration figure is therefore computed on a subset whose membership is partly determined by wording choices that are rarely reported.

Much of that missingness belongs to the extraction rule rather than to the model. Two of the requests we added ask for an integer between 0 and 100, and the parser used throughout accepts a bare integer only when it constitutes the whole reply, so under it those two score as missing almost everywhere. Yet roughly three quarters of the discarded responses contain a perfectly readable number, in forms such as `Likelihood: 80`. Reading the first integer that is not part of a decimal, and dividing by the scale the prompt itself specifies, recovers a usable value in the large majority of cases (Appendix D). Whether a confidence value is recorded therefore depends on whether the scale the request asks for matches the pattern the scorer expects, a pairing that reported calibration numbers do not state.

4.3 The signal is not stable under paraphrase of the request

This repeated-elicitation design holds constant everything except the wording of the confidence question. The item, the model and the model's own answer are fixed, and the answer is spliced back into the context before the suffix is appended, so any disagreement between two phrasings is attributable to the phrasing.

Comparing the two phrasings from prior work, both of which parse on close to all items, agreement is low (Table 6). Across 94 settings with a median of 498 shared items each, the mean Pearson correlation is 0.115, falling below 0.30 in 81 of them; the mean Spearman correlation is 0.105, which matters because the stated values are heavily quantised. One further setting is dropped because one phrasing is constant there, leaving the correlation undefined.

A low correlation on its own is not decisive, because the signal has small variance and range restriction would depress any correlation computed on it. We therefore compare the variation that rewording introduces against the variation across questions. The final column of Table 6 compares the standard deviation of the phrasing-induced difference against that of the signal itself: the ratio exceeds one in 73 of the 94 and averages 1.14, so rewording moves the stated value more than the value varies across items. Llama-3.2-3B makes the point without the ratio, having a confidence distribution three to six times wider than the others, the condition under which range restriction would predict the highest correlation, and the lowest correlation at 0.024 (Figure 3, Appendix D).

A second objection is that this result could be specific to the particular pair of formulations. Repeating the analysis on the three added later, which parse on a median of 93 to 99 percent of items once the scorer's scale assumption is corrected, gives a mean correlation of 0.069 across 208 settings with at least 100 shared items, below 0.30 in 191 of them, against 0.079 over the 305 settings formed by all six pairs of the first four (Appendix D). Three phrasings on which the measurement is nearly always available agree with one another no better than the two from prior work do. These three were generated in bfloat16 and the earlier four in float16, so the two blocks are reported separately and never pooled.

A third objection is more fundamental: no two wordings are ever exactly the same request, so disagreement between them might reflect a genuine difference in what was asked. To avoid relying on semantic equivalence at all, we ask the complementary question. Instead of how likely the answer is to be right, we ask how likely it is to be wrong. If the reported numbers behave like probabilities, the two should sum to one on each item. They do not: the item-level agreement between the complement of the stated error probability and the stated probability of correctness averages 0.112. Four of the five models are close to coherent when averaged over many questions while agreeing hardly at all on individual ones, which is what a signal that is roughly right on average and uninformative about the particular case looks like. The fifth departs in aggregate too, stating on average that its answer is 84% likely to be right and 64% likely to be wrong (Appendix D).

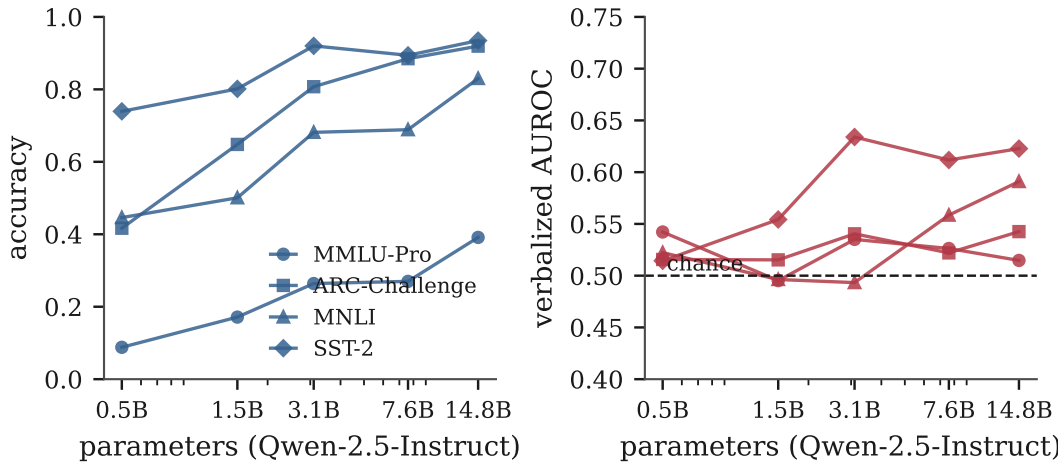

Figure 2: The Qwen-2.5-Instruct scaling series, 0.5B to 32.8B at fixed recipe and tokenizer. Left: accuracy. Right: AUROC for both signals on the same settings, token probability in blue and the verbalized report in red. Accuracy and token probability rise on every benchmark while the verbalized signal stays near chance, so the distance between them grows with scale.

4.4 Neither scale nor instruction tuning changes the picture

Better models do not bring better verbal confidence. The scaling series holds family, tokenizer and training recipe fixed across a sixty-six-fold range in parameters (Figure 2; Table 3). Task accuracy improves steeply on all five benchmarks, rising more than fourfold on MMLU-Pro. So does the informativeness of token probability. The verbalized report improves as well, but roughly three times more slowly, so the distance between what the model encodes and what it can state widens rather than closes over the range. On MMLU-Pro, the hardest of the five, the verbalized signal does not improve at all: its correlation with size is negative and it stays within a few points of chance while token probability climbs from 0.556 to 0.793.

Instruction tuning changes the format, not the content. Across the same 125 settings on each side, the five base models have a mean parse rate of 0.543 and their instruction-tuned counterparts 0.793, while verbal AUROC over the settings clearing the threshold is 0.500 and 0.522. The individual pairs make the dissociation clearer: Gemma-3-4B moves from a mean parse rate of 0.032 to 0.993 while verbal AUROC remains at 0.502. Instruction tuning appears to confer the ability to emit a confidence-shaped output rather than the ability to estimate correctness. Settings in which the base models clear the threshold are fewer, 55 against 94, and are selected by that threshold, so the base-side estimate is measured only where base models happen to parse.

Explicit confidence training is different. Under the same procedure the ConfTuner checkpoint of Section 3.3 reaches a verbal AUROC of 0.605 against 0.517 for its untuned base, the modal value falls from 79.8% to 54.9% of responses, and the number of distinct values roughly doubles. The measurement registers a trained verbal signal. The gain is uneven across our benchmarks (Table 4), from 0.512 to 0.725 on SST-2 but 0.533 to 0.540 on MMLU-Pro, which is expected given that these are not the benchmarks the checkpoint was evaluated on. Even so, the tuned model's verbal signal remains below its own token probability, 0.605 against 0.707.

4.5 Consequences for reported comparisons and for downstream use

Comparative claims. Ranking the five primary models by verbal calibration error and then changing only the elicitation phrasing, the model that appears best calibrated changes in 4 of 14 benchmark and variant comparisons, counting every comparison in which at least three models yield a defined verbal ECE under both phrasings; mean Kendall correlation between the two orderings is 0.471 (Figure 4 in Appendix D). Including the two excluded formulations gives 9 of 19 and 0.537.

A statement that one model is better calibrated than another, when the calibration is verbal, may therefore be reporting a property of the prompt.

Selective prediction. The most commonly cited use for a confidence signal is deciding when to trust an answer. Retaining the most confident fraction of items and measuring accuracy on what remains, averaged over the 100 settings where both signals clear the parse threshold, token probability yields roughly seven times the gain at every operating point. At 50% coverage the verbalized signal improves accuracy by 0.010 against 0.072, and it produces no improvement at all in 35 of them. An abstention or routing policy built on this signal recovers little of what is available from a quantity available to anyone with logit access to the same model.

5 Discussion

5.1 What the results do and do not establish

It helps to place these results beside the other quantities an evaluation reports. Decomposing each metric's variation within a benchmark into model, prompt and interaction terms (Appendix E), accuracy is majority attributable to the model, which is the property a benchmark is supposed to have. Verbalized AUROC is the opposite case: most of its variation sits in the interaction, the pattern expected when there is no stable quantity to attribute.

The properties we measure are ordered. A signal that cannot be obtained is not available for calibration; one that changes when the request is reworded, or that contradicts its own answer to the complementary question, has no calibration property belonging to the model; and one that does not order examples by correctness cannot support a decision about which answers to trust. Verbalized confidence fails at each level, and the failures are not independent: a near-constant signal will correlate poorly with itself under any perturbation, and its calibration error is largely determined by accuracy. Expected calibration error did not reveal this because it is defined for any confidence input, including a constant one. A model reporting 0.9 on every item has a well defined ECE equal to the distance between 0.9 and its accuracy, which can be small when accuracy sits near 0.9, and nothing in the metric signals that the input carried no information. The check we propose is inexpensive: AUROC on the same responses requires no additional inference, is threshold-free, and is invariant to the monotone rescalings calibration methods apply.

The positive control constrains the interpretation. Because a checkpoint trained for this purpose shows a higher AUROC and a broader value distribution under the same measurement, the null results are properties of the models we measured rather than of our instrument, and they are not evidence that verbalized confidence cannot work. Nor do we conclude that token probability is well calibrated: its ECE is lower, 0.180 against 0.307, but calibration was not our subject. What we show is narrower, that token probability orders examples by correctness usefully and verbalized confidence does not, on the same responses.

5.2 Recommendations

We make four, each computable from quantities an evaluation already produces. *Report an informativeness check alongside any verbal calibration number*; AUROC for the signal predicting correctness is the natural choice, and where it is close to 0.5 the accompanying ECE should not be read as evidence about the model's self-knowledge. *Report the parse rate for each setting, with the inclusion rule and the extraction pattern*: a verbal calibration mean is conditional on the subset where a value could be extracted, and that subset is selected both by the elicitation wording and by the scorer, so a mean computed this way is not comparable across papers using different wordings or different extraction code. *Replicate any verbal confidence claim under at least two ways of asking and report the agreement between them*; a single-wording result should not be read as a property of the model until that agreement is known to be high. *Prefer token probability where it is available*: on every measure we examined it was the better signal. The practical arguments for verbalized confidence are real, since it survives an API boundary and applies to free-form output, but they should be weighed against a measured difference in signal quality rather than assumed to be costless.

6 LIMITATIONS

Our benchmarks are English, single-label and closed-set. Confidence in open-ended generation is a different measurement problem, since correctness itself must then be estimated, and nothing here transfers to it directly.

The scaling series spans 0.5B to 32.8B. We have no evidence about larger models, and the scaling result establishes only that the verbalized signal improves more slowly than the token signal over the range we tested under one training recipe; extrapolating either trend beyond six points is not warranted.

Our seven ways of requesting confidence bound rather than exhaust the space. The three added last parse well and reproduce the low agreement found between the two taken from prior work, so the result no longer rests on a subset selected by parse failure; those three ran in bfloat16 and the earlier four in float16 on a device without bfloat16 support, so pairs crossing the two runs confound wording with numerical precision and we exclude them (Table 5). A systematic sweep over families of request, including chain-of-thought self-assessment and categorical scales, would be a useful extension.

The base and instruct comparison rests on five pairs, and the settings that enter on the base side are selected by the same parse-rate threshold that base models often fail, so the estimate for the base models is measured only where they produce parseable output. The positive control is a single checkpoint from a single method, evaluated outside the benchmarks it was developed on; it establishes that our pipeline detects a trained signal but does not characterise how well trained verbal confidence performs in general.

All inference used one engine with greedy decoding, and we did not test whether the token-probability comparisons are stable across engines or numerical precisions. The variance attribution is descriptive: its model term depends on how heterogeneous the model set is, so we read the ordering across metrics rather than the magnitudes.

7 CONCLUSION

Asking a language model how confident it is produces a number, and that number is widely reported as a calibration result. We tested whether it is a measurement. In fourteen instruction-tuned open-weight models it orders answers by correctness little better than chance, it does not agree with itself when the request is reworded or with its own answer to the complementary question, and it frequently cannot be obtained at all. Scale does not resolve this, and instruction tuning makes the number easier to obtain without making it more informative. A model trained for the purpose behaves differently under the same procedure, which is how we know the instrument works, and none of this is visible in a calibration error alone.

AI USE STATEMENT

Generative AI tools were used for writing and coding assistance. All research ideas, experimental work, analysis, and conclusions were developed and evaluated by the authors. The final manuscript reflects the authors' original work, and the authors take full responsibility for its content, results, and claims.

REPRODUCIBILITY STATEMENT

We provide the prompts, confidence elicitation formats, parsing rules, and analysis procedures used in this work in the appendices. The released materials include the analysis code and model outputs used to produce the reported results, figures, and tables. This allows the results to be reproduced and the outputs to be reanalyzed using alternative parsing or answer-extraction rules.

ETHICS STATEMENT

This work uses publicly available model checkpoints and benchmarks and involves no human subjects, personal data, or new data collection. Our results are intended to improve the evaluation of

language-model confidence. We do not make claims about the suitability of any particular model for specific real-world uses.

REFERENCES

Norah Alzahrani, Hisham Abdullah Alyahya, Yazeed Alnumay, Sultan Alrashed, Shaykhah Alsubaie, Yusef Almushaykeh, Faisal Mirza, Nouf Alotaibi, Nora Altwairesh, Areeb Alowisheq, M Saiful Bari, and Haidar Khan. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards, 2024. URL https://arxiv.org/abs/2402.01781.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, Anthony DiPofi, Julen Etxaniz, Benjamin Fattori, Jessica Zosa Forde, Charles Foster, Jeffrey Hsu, Mimansa Jaiswal, Wilson Y. Lee, Haonan Li, Charles Lovering, Niklas Muennighoff, Ellie Pavlick, Jason Phang, Aviya Skowron, Samson Tan, Xiangru Tang, Kevin A. Wang, Genta Indra Winata, François Yvon, and Andy Zou. Lessons from the trenches on reproducible evaluation of language models. 2026. URL https://arxiv.org/abs/2405.14782.

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

Mehul Damani, Isha Puri, Stewart Slocum, Idan Shenfeld, Leshem Choshen, Yoon Kim, and Jacob Andreas. Beyond binary rewards: Training lms to reason about their uncertainty, 2026. URL https://arxiv.org/abs/2507.16806.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 1321–1330, 2017.

Abigail Z. Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 375–385. ACM, March 2021. doi: 10.1145/3442188.3445901. URL http://dx.doi.org/10.1145/3442188.3445901.

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

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th ACM Symposium on Operating Systems Principles (SOSP)*, pp. 611–626, 2023.

Emeline Laborie and Andrew Callan-Jones. Power-law decay of force on cell membrane tethers reflects long-ranged relaxation of membrane tension, 2025. URL https://arxiv.org/abs/2502.05601.

Yibo Li, Miao Xiong, Jiaying Wu, and Bryan Hooi. Conftuner: Training large language models to express their confidence verbally, 2025. URL https://arxiv.org/abs/2508.18847.

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

Zhengzhao Ma, Xueru Wen, Boxi Cao, Yaojie Lu, Hongyu Lin, Jinglin Yang, Min He, Xianpei Han, and Le Sun. Decoupling reasoning and confidence: Resurrecting calibration in reinforcement learning from verifiable rewards, 2026. URL https://arxiv.org/abs/2603.09117.

Meta AI. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, Mei Gao, Min Gao, Amit Garg, Abhishek Goswami, Junheng Hao, Amr Hendy, Yuxuan Hu, Xin Jin, Mahmoud Khademi, Dongwoo Kim, Young Jin Kim, Gina Lee, Jinyu Li, Yunsheng Li, Chen Liang, Xihui Lin, Zeqi Lin, Mengchen Liu, Yang Liu, Gilsinia Lopez, Chong Luo, Piyush Madan, Vadim Mazalov, Arindam Mitra, Ali Mousavi, Anh Nguyen, Jing Pan, Daniel Perez-Becker, Jacob Platin, Thomas Portet, Kai Qiu, Bo Ren, Liliang Ren, Sambuddha Roy, Ning Shang, Yelong Shen, Saksham Singhal, Subhojit Som, Xia Song, Tetyana Sych, Praneetha Vaddamanu, Shuohang Wang, Yiming Wang, Zhenghao Wang, Haibin Wu, Haoran Xu, Weijian Xu, Yifan Yang, Ziyi Yang, Donghan Yu, Ishmam Zabir, Jianwen Zhang, Li Lyna Zhang, Yunan Zhang, and Xiren Zhou. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras, 2025. URL https://arxiv.org/abs/2503.01743.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. State of what art? a call for multi-prompt llm evaluation, 2024. URL https://arxiv.org/abs/2401.00595.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pp. 2901–2907, 2015.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. Ai and the everything in the whole wide world benchmark. 2021. URL https://arxiv.org/abs/2111.15366.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1631–1642, 2013.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. 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 (EMNLP)*, pp. 5433–5442, 2023.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024.

Albert Webson and Ellie Pavlick. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 2300–2344, 2022.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 1112–1122, 2018.

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

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems 28 (NeurIPS)*, pp. 649–657, 2015.