
One small model’s confidence mostly sits on one value, and asking after the answer separates right from wrong no better

Humanity First Research

Abstract

Does moving a confidence request after the answer improve separation of correct from incorrect responses? We tested Qwen2.5 7B on 400 questions over five seeds using hundred-point, five-step, and forced-choice reports. Pooled order movements were 0.0829 points and -0.0405 steps; paired-seed 95% intervals were -0.675 to 0.796 and -0.186 to 0.0823. The prediction was not supported, although modest effects remain possible. Modal share, the share of replies at the most common value, rose from 0.883 to 0.930 for hundred-point reports and from 0.927 to 0.968 for five-step reports. Forced choice had modal share 0.371 and area under the ROC curve (AUROC) 0.606, but cannot be elicited before an answer. Exploratorily, answer-span probability exceeded stated-value AUROC under both numeric forms, by 0.111 and 0.108, but not detectably under forced choice.

Hypothesis

A model asked how sure it is can say so in words, and the number it states need not match the probability its own weights assign to the answer. Numbers given that way have been reported to track accuracy more closely than those internal probabilities do (Tian et al. 2023). If the spoken number carries information, when it is asked for ought to matter.

Recent work builds that ordering into the method, generating an answer first and estimating confidence against the question and answer once both are fixed (Li et al. 2026). The reasoning behind the choice is plain. A number stated before any answer exists has nothing to condition on but the question, so it can express at most how hard the question looks. One stated afterwards can condition on what was actually produced. Should that difference matter, moving the request ought to widen the gap between the sureness stated on right answers and on wrong ones.

That is the prediction under test: asked after the answer rather than before, this model’s stated sureness will separate its right answers from its wrong ones more widely than it does when asked first. Two things could prevent it. Most evidence for spoken sureness comes from large models, which are reported to carry high calibration error and to be overconfident most of the time (Groot and Valdenegro-Toro 2024), and a model that declares itself nearly certain of everything leaves no gap for either order to widen. A model that ignored where the request sat and answered in whatever order it preferred would collapse the two conditions into one, which is a failure of the manipulation rather than of the idea.

Setup

We evaluated Qwen2.5 7B (Ollama tag qwen2.5:7b), a 7.6-billion-parameter instruction-tuned model quantised to Q4_K_M, at temperature 0.7 and top_p 0.95 with a 64-token cap (16 for the forced choice). Five seeds, 7001 through 7005, covered 400 single-word factual questions and produced 12,000 rows. The study had two runs: the first compared

hundred-point confidence before and after an answer across 4,000 generations in 270 seconds on one consumer GPU; the second compared labelled steps before and after and added forced choice in 8,000 new generations and 325 seconds, reusing the first run’s hundred-point rows as its baseline. The frozen question file is preserved with the run’s data. The shipped builder regenerates it byte for byte from geonamescache 3.0.2, derived from GeoNames, and periodictable 2.1.0; this SHA-256 verifies any copy.

`e72f5e5f35fc8782f095c901bbebf1be029ae550e0fbc3d43975f3d56adfd2ae`

Hundred-point and labelled-step prompts used identical questions. The orders differ in the first instruction line, prospective (“Before answering, state how sure you are that you will get it right, then answer.”) versus retrospective (“Answer, then state how sure you are that you got it right.”), and in the order of the two reply lines. Line labels, scale, and the two-line constraint are unchanged.

Predefined readers extracted and normalised the first labelled answer line, matched the gold or an alias as a whole word, and read the first confidence number or one of five ordered labels. The serving interface returned a log probability for every generated token; tokens whose byte ranges overlapped the extracted answer span were collected, and the exponential of their mean log probability was stored beside the answer as the mean token probability (the geometric mean of the token probabilities). Forced choice paired adjacent questions in a cycle, presented both pair orders, and assigned each question the share of four comparisons won per seed. The registered order metric was the correct-minus-incorrect difference in mean stated value, predicting an after-minus-before increase. No statistical test was registered; the analysis used paired t tests across the five seeds for the order movements, Welch tests for each order’s distance, a paired t test over questions for the accuracy difference, and percentile bootstraps over questions for intervals and area comparisons.

Kadavath et al. scored an already-proposed answer with $P(\text{True})$ and trained a before-answer predictor, $P(\text{IK})$, in large models (Kadavath et al. 2022); ORCE builds answer-first confidence estimation into training (Li et al. 2026). Verbalised confidence is often overconfident and piled on a few values (Xiong et al. 2024; Groot and Valdenegro-Toro 2024), and can be trained to be calibrated (Lin et al. 2022); our delta is to move only the position of an untrained request in one small model.

Results

Both manipulations passed: confidence followed the answer in 0 of before-order replies and 1.0 after-order, while labelled-step compliance rose from 0 to 0.996. Readers discarded no before-order and two after-order hundred-point replies, plus none of 4,000 forced-choice presentations.

In the first run, hundred-point distances between correct and incorrect replies were 2.74 points before and 2.83 after; the separate first-run tests gave $p = 5.92641\text{e-}06$ on an interval of 1.5665 to 3.9213 and $p = 3.01046\text{e-}06$ on 1.6504 to 4.0031. The pooled movement was 0.0829 points against a 0.592 seed spread, while a paired test over five seed movements gave $t(4) = 0.228$, $p = 0.83$, and -0.675 to 0.796 ; question resampling gave -3.08 to 3.28 at $p = 0.97$. In the second run, labelled-step distances were 0.195 before and 0.154 after, a movement of -0.0405 steps against a 0.108 seed spread, with $t(4) = -1.07$, $p = 0.34$, and -0.186 to 0.0823 ; question resampling gave -0.165 to 0.0833 at $p = 0.50$.

The prediction was not supported: both movements were smaller than their seed spread, the registered outcome that weakens the idea. No minimum effect was registered, and paired-seed intervals allow increases of 0.80 points and 0.082 steps. From the reported seed spreads, a conventional calculation gives five seeds 80% power only for movements of about 1.0 point and 0.18 steps.

The distributions explain the limited movement: hundred-point modal shares were 0.883 and 0.930 across 9 and 5 values, labelled-step shares were 0.927 and 0.968 across 4 values each, and forced choice used all 5 values with 0.371 at its mode. Pooled correct-minus-incorrect distances were 2.78 points (1.95 to 3.62), 0.171 steps (0.130 to 0.213), and 0.101 in win share (0.0749 to 0.127).

Stated-value area under the ROC curve (AUROC) was 0.567, 0.548, and 0.606; relative to hundred-point reports, labelled steps changed AUROC by -0.0186 (-0.0408 to 0.0017, $p = 0.082$) and forced choice by 0.0391 (-0.0129 to 0.0918, $p = 0.174$), with neither difference detected. The registered primary analyses were the hundred-point order test and labelled-versus-hundred-point AUROC; the design also prespecified the labelled-step order contrast, but none was significant. Of the 11 reported tests, the five with unadjusted p below 0.05 all survive Holm correction: the two within-order distances and three span-probability comparisons.

The span-probability comparison is exploratory, not a registered primary test; its AUROC was 0.678, 0.656, and 0.656. Over the narrower set of first-run replies carrying a span probability, the hundred-point advantage was 0.1106 on 0.0483 to 0.1769, at $p = 0.0005$. The second run recomputed every undiscarded reply: advantages were 0.111 for hundred-point reports (0.0484 to 0.177, $p = 0.001$), 0.108 for labelled steps (0.0506 to 0.169, $p = 0.001$), and 0.0496 for forced choice (-0.0282 to 0.124, $p = 0.19$). Tian et al. found stated confidence better calibrated than token probabilities in larger feedback-tuned models (Tian et al. 2023), whereas Xiong et al. found white-box probabilities narrowly ahead of verbalised confidence, matching our ordering (Xiong et al. 2024).

Choices picked the first-presented option at a rate of 0.494 with seed spread 0.0073, close to the 0.5 expected without a position preference, while hundred-point accuracy moved from 0.767 to 0.748, a -1.90-point difference with $t = -1.72$, $p = 0.085$, and -4.07 to 0.266. Labelled-step accuracy moved from 0.762 to 0.736, a difference the runs did not test.

Figures

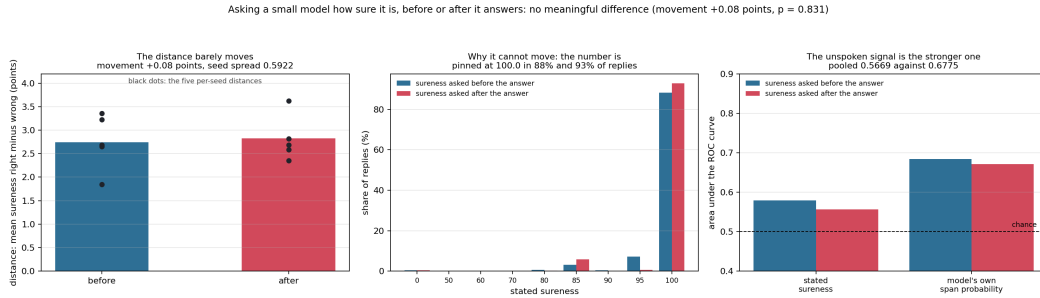


Figure 1: Figure 1. In the first run, moving the confidence request changed the pooled correct-minus-incorrect distance by 0.0829 points, smaller than the 0.592 standard deviation across the five seed movements.

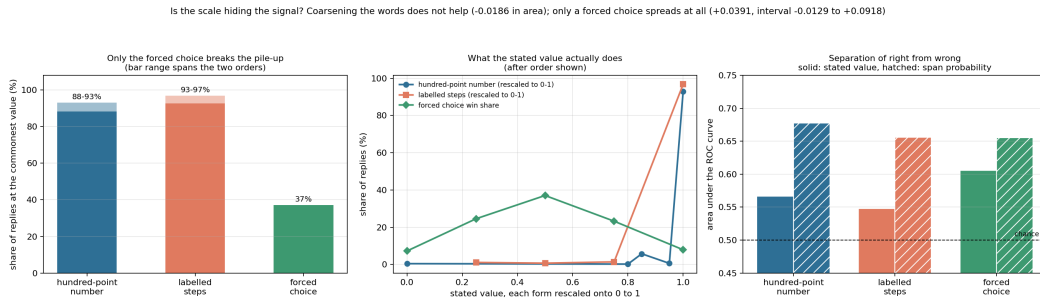


Figure 2: Figure 2. In the second run, five labelled steps increased the before/after modal shares to 0.927 and 0.968, compared with 0.883 and 0.930 for the first run's hundred-point baseline.

Limitations

One model at one size, quantised and served locally, stands behind everything here, and the three forms are three among many. The follow-up changed the vocabulary and left the

model alone, so nothing in either run speaks to a larger model, a differently tuned one, or one trained to report its confidence well.

The item set was constructed offline rather than taken from a published benchmark, so its scope is still limited by the source tables and question templates. The shipped builder now supplies the missing reproducibility path: a runner-recorded rebuild wrote all 400 questions, produced a file byte-identical to the preserved copy, and returned SHA-256 `e72f5e5f35fc8782f095c901bbecf1be029ae550e0fbc3d43975f3d56adfd2ae`. This verifies that the frozen set can be regenerated from the shipped code. It does not make a synthetic, template-built set representative of factual questions more broadly.

What the follow-up set out to do was find a form with room to move, and it found one. That is also where it stops. The form that spreads judges answers that already exist and cannot be asked before one does, so the question the paper began with is still untested on any form with room in it. Of the two nulls the labelled-step one is more informative, because the vocabulary changed underneath it, but it was measured where more than nine replies in ten sit on a single step.

Both cross-form comparisons carry intervals that include zero. Neither the claim that the labelled steps are worse than the number nor the claim that the forced choice is better is established by this work. What is established is narrower: the labelled steps did not break the pile-up and the forced choice did.

The forced choice is also a coarse instrument. Each question gets four comparisons in a seed, so its value takes one of five levels, and a value built from four binary judgements is a noisy estimate of whatever underlies it. More pairings for each question would sharpen it, and would settle whether its apparent parity with the model’s own probability is real or the noise of a small denominator.

Because the stated value is nearly constant, the distances rest on a thin tail. Almost the whole signal comes from the minority of replies that said something other than the commonest value, which the intervals reflect but which is easy to forget when the figures are printed to four places.

Under the hundred-point number, accuracy differs between the two orders by 1.9 points, which is not significant and not nothing. The same comparison under the labelled steps gives 0.7615 and 0.7355, a wider gap the runs leave untested. Either way the right and wrong sets are not quite the same replies in the two arms. Against a null it is harmless, since a reshuffle of that size cannot conceal an effect as large as the one predicted, but any future run reporting a positive effect would have to deal with it.

Two controls were registered and then waived, and the first bears on the contrast the whole paper rests on. The wordings of the two sureness requests were never matched, and the instruction blocks differ in their one framing sentence as well as in the order of the two requested lines, so a position effect and a phrasing effect are confounded throughout. That confound cannot explain what was found here, because nothing was found to explain, but it would matter to any future run that did find something. The second waiver leaves the hundred-point baseline in a different session from the forms compared against it, and both cross-form differences are small enough that session drift cannot simply be dismissed as an explanation.

Last, the comparison with the model’s own probability was a primary estimate in neither run. Unregistered in the first, it rests in the second on a recording the design registered as a control. It is measured on the answer span rather than the whole reply, and a mean over the covering tokens is one of several defensible ways to summarise that span. A run built to test it would fix that choice in advance rather than inherit it from a rule written to record something else.

Reproducibility

This appendix is generated mechanically from the run’s recorded artifacts so that every experimental cell, its sample size, its seeds, and its compute cost are inspectable without

re-running the job or asking the authors. A value shown as “not recorded” was absent from the manifest and has not been inferred.

The run comprises 40 recorded experiments. Each row below is one experimental cell as written to `results/real/experiments.json`.

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyper- parameters	GPU min- utes
after__order	qwen2.5:7b	frozen question file (questions.- jsonl): 400 short factual questions with single-word gold answers, mixed difficulty	inference	2000	not recorded	0 steps; params=7.6B	0.00
before__order	qwen2.5:7b	frozen question file (questions.- jsonl): 400 short factual questions with single-word gold answers, mixed difficulty	inference	2000	not recorded	0 steps; params=7.6B	0.00
main__- after__- order__- seed7001	qwen2.5:7b	frozen question file (questions.- jsonl): 400 short factual questions with single-word gold answers, mixed difficulty	inference	400	7001	0 steps; params=7.6B	0.45

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyper- parameters	GPU min- utes
main__- after__- order__- seed7002	qwen2.5:7b	frozen question file (questions.- jsonl): 400 short factual questions with single-word gold answers, mixed difficulty	inference	400	7002	0 steps; params=7.6B	0.45
main__- after__- order__- seed7003	qwen2.5:7b	frozen question file (questions.- jsonl): 400 short factual questions with single-word gold answers, mixed difficulty	inference	400	7003	0 steps; params=7.6B	0.45
main__- after__- order__- seed7004	qwen2.5:7b	frozen question file (questions.- jsonl): 400 short factual questions with single-word gold answers, mixed difficulty	inference	400	7004	0 steps; params=7.6B	0.45
main__- after__- order__- seed7005	qwen2.5:7b	frozen question file (questions.- jsonl): 400 short factual questions with single-word gold answers, mixed difficulty	inference	400	7005	0 steps; params=7.6B	0.45

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyper- parameters	GPU min- utes
main__- before__- order__- seed7001	qwen2.5:7b	frozen question file (questions.- jsonl): 400 short factual questions with single-word gold answers, mixed difficulty	inference	400	7001	0 steps; params=7.6B	0.45
main__- before__- order__- seed7002	qwen2.5:7b	frozen question file (questions.- jsonl): 400 short factual questions with single-word gold answers, mixed difficulty	inference	400	7002	0 steps; params=7.6B	0.45
main__- before__- order__- seed7003	qwen2.5:7b	frozen question file (questions.- jsonl): 400 short factual questions with single-word gold answers, mixed difficulty	inference	400	7003	0 steps; params=7.6B	0.45
main__- before__- order__- seed7004	qwen2.5:7b	frozen question file (questions.- jsonl): 400 short factual questions with single-word gold answers, mixed difficulty	inference	400	7004	0 steps; params=7.6B	0.46

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyper- parameters	GPU min- utes
main__- before__- order__- seed7005	qwen2.5:7b	frozen question file (questions.- jsonl): 400 short factual questions with single-word gold answers, mixed difficulty	inference	400	7005	0 steps; params=7.6B	0.46
forced__- choice	qwen2.5:7b	the first run's frozen question file (questions.- jsonl): 400 short factual questions with single-word gold answers	inference	4000	not recorded	0 steps; params=7.6B	0.00
labelled__- steps	qwen2.5:7b	the first run's frozen question file (questions.- jsonl): 400 short factual questions with single-word gold answers	inference	4000	not recorded	0 steps; params=7.6B	0.00
hundred__- point	qwen2.5:7b	the first run's frozen question file (questions.- jsonl): 400 short factual questions with single-word gold answers	inference	4000	not recorded	0 steps; params=7.6B	0.00

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyper- parameters	GPU min- utes
choice__- seed7001	qwen2.5:7b	the first run's frozen question file (questions.- jsonl): 400 short factual questions with single-word gold answers	inference	800	7001	0 steps; params=7.6B	0.33
choice__- seed7002	qwen2.5:7b	the first run's frozen question file (questions.- jsonl): 400 short factual questions with single-word gold answers	inference	800	7002	0 steps; params=7.6B	0.33
choice__- seed7003	qwen2.5:7b	the first run's frozen question file (questions.- jsonl): 400 short factual questions with single-word gold answers	inference	800	7003	0 steps; params=7.6B	0.33
choice__- seed7004	qwen2.5:7b	the first run's frozen question file (questions.- jsonl): 400 short factual questions with single-word gold answers	inference	800	7004	0 steps; params=7.6B	0.33

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyper- parameters	GPU min- utes
choice__- seed7005	qwen2.5:7b	the first run's frozen question file (questions.- jsonl): 400 short factual questions with single-word gold answers	inference	800	7005	0 steps; params=7.6B	0.33
labelled__- after__- order__- seed7001	qwen2.5:7b	the first run's frozen question file (questions.- jsonl): 400 short factual questions with single-word gold answers	inference	400	7001	0 steps; params=7.6B	0.37
labelled__- after__- order__- seed7002	qwen2.5:7b	the first run's frozen question file (questions.- jsonl): 400 short factual questions with single-word gold answers	inference	400	7002	0 steps; params=7.6B	0.36
labelled__- after__- order__- seed7003	qwen2.5:7b	the first run's frozen question file (questions.- jsonl): 400 short factual questions with single-word gold answers	inference	400	7003	0 steps; params=7.6B	0.37

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyper- parameters	GPU min- utes
labelled__- after__- order__- seed7004	qwen2.5:7b	the first run's frozen question file (questions.- jsonl): 400 short factual questions with single-word gold answers	inference	400	7004	0 steps; params=7.6B	0.37
labelled__- after__- order__- seed7005	qwen2.5:7b	the first run's frozen question file (questions.- jsonl): 400 short factual questions with single-word gold answers	inference	400	7005	0 steps; params=7.6B	0.37
labelled__- before__- order__- seed7001	qwen2.5:7b	the first run's frozen question file (questions.- jsonl): 400 short factual questions with single-word gold answers	inference	400	7001	0 steps; params=7.6B	0.39
labelled__- before__- order__- seed7002	qwen2.5:7b	the first run's frozen question file (questions.- jsonl): 400 short factual questions with single-word gold answers	inference	400	7002	0 steps; params=7.6B	0.38

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyper- parameters	GPU min- utes
labelled__- before__- order__- seed7003	qwen2.5:7b	the first run's frozen question file (questions.- jsonl): 400 short factual questions with single-word gold answers	inference	400	7003	0 steps; params=7.6B	0.39
labelled__- before__- order__- seed7004	qwen2.5:7b	the first run's frozen question file (questions.- jsonl): 400 short factual questions with single-word gold answers	inference	400	7004	0 steps; params=7.6B	0.39
labelled__- before__- order__- seed7005	qwen2.5:7b	the first run's frozen question file (questions.- jsonl): 400 short factual questions with single-word gold answers	inference	400	7005	0 steps; params=7.6B	0.38
reused__- b4623192__- main__- after__- order__- seed7001	qwen2.5:7b	the first run's frozen question file (questions.- jsonl): 400 short factual questions with single-word gold answers	inference	400	7001	0 steps; params=7.6B	0.00

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyper- parameters	GPU min- utes
reused_- b4623192_- main_- after_- order_- seed7002	qwen2.5:7b	the first run's frozen question file (questions.- jsonl): 400 short factual questions with single-word gold answers	inference	400	7002	0 steps; params=7.6B	0.00
reused_- b4623192_- main_- after_- order_- seed7003	qwen2.5:7b	the first run's frozen question file (questions.- jsonl): 400 short factual questions with single-word gold answers	inference	400	7003	0 steps; params=7.6B	0.00
reused_- b4623192_- main_- after_- order_- seed7004	qwen2.5:7b	the first run's frozen question file (questions.- jsonl): 400 short factual questions with single-word gold answers	inference	400	7004	0 steps; params=7.6B	0.00
reused_- b4623192_- main_- after_- order_- seed7005	qwen2.5:7b	the first run's frozen question file (questions.- jsonl): 400 short factual questions with single-word gold answers	inference	400	7005	0 steps; params=7.6B	0.00

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyper- parameters	GPU min- utes
reused__- b4623192__- main__- before__- order__- seed7001	qwen2.5:7b	the first run's frozen question file (questions.- jsonl): 400 short factual questions with single-word gold answers	inference	400	7001	0 steps; params=7.6B	0.00
reused__- b4623192__- main__- before__- order__- seed7002	qwen2.5:7b	the first run's frozen question file (questions.- jsonl): 400 short factual questions with single-word gold answers	inference	400	7002	0 steps; params=7.6B	0.00
reused__- b4623192__- main__- before__- order__- seed7003	qwen2.5:7b	the first run's frozen question file (questions.- jsonl): 400 short factual questions with single-word gold answers	inference	400	7003	0 steps; params=7.6B	0.00
reused__- b4623192__- main__- before__- order__- seed7004	qwen2.5:7b	the first run's frozen question file (questions.- jsonl): 400 short factual questions with single-word gold answers	inference	400	7004	0 steps; params=7.6B	0.00

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyper- parameters	GPU min- utes
reused_- b4623192_- main_- before_- order_- seed7005	qwen2.5:7b	the first run’s frozen question file (questions.- jsonl): 400 short factual questions with single-word gold answers	inference	400	7005	0 steps; params=7.6B	0.00

Seed policy. The distinct RNG seeds recorded across the manifest are 7001, 7002, 7003, 7004, 7005. Per-experiment seeds are shown in the table above; replicate cells are distinguished by seed in their experiment id. 5 experiment(s) carry no recorded seed (single-shot supervisor/ceiling fits or eval passes) and are marked “not recorded”.

Cross-validation. No cross-validation fold fields are recorded in the manifest.

Statistical tests. Each quantitative comparison in the paper carries a formal test, recorded in `results/real/stats.json`.

Claim	Test	Statistic	p	95% CI	n	Seeds
Primary: asking for the sureness after the answer rather than before it does not widen the distance between the sureness stated on right answers and on wrong ones.	paired t-test across the five seeds on the per-seed distance difference, with a percentile bootstrap over questions for the interval	0.0829	0.83	[-3.08, 3.28]	400	5
In the before order the stated sureness is higher on right answers than on wrong ones, by a small amount.	Welch two-sample t-test on right minus wrong stated sureness	2.74	0.0000059	[1.57, 3.92]	2000	5
In the after order the stated sureness is higher on right answers than on wrong ones, by a similarly small amount.	Welch two-sample t-test on right minus wrong stated sureness	2.83	0.000003	[1.65, 4]	1998	5

Claim	Test	Statistic	p	95% CI	n	Seeds
Accuracy is slightly lower when the sureness is requested after the answer, which matters because accuracy decides which replies the distance is measured over.	paired t-test over questions on the per-question accuracy difference averaged across the five seeds	-1.9	0.085	[-4.07, 0.266]	400	5
Secondary: the model's own probability for the answer span it produced separates right from wrong answers far better than the sureness it states in words.	difference of two areas under the ROC curve, with a percentile bootstrap over questions resampling both scores together	0.111	0.0005	[0.0483, 0.177]	3997	5
Primary: swapping the hundred-point number for a few labelled steps does not separate right answers from wrong ones any better; if anything it is slightly worse.	difference of two areas under the ROC curve, percentile bootstrap over questions resampling both forms together	-0.0186	0.082	[-0.0408, 0.0017]	400	5
The forced choice separates right from wrong better than the hundred-point number, but the interval over questions still includes no difference.	difference of two areas under the ROC curve, percentile bootstrap over questions resampling both forms together	0.0391	0.17	[-0.0129, 0.0918]	400	5
The order contrast the hypothesis asked for, run again under the labelled steps, is null: asking after the answer does not widen the distance.	paired t-test across the five seeds on the per-seed distance difference, with a percentile bootstrap over questions for the interval	-0.0405	0.34	[-0.165, 0.0833]	400	5

Claim	Test	Statistic	p	95% CI	n	Seeds
Within the hundred-point form the model’s own span probability separates right from wrong better than the value it states.	difference of two areas under the ROC curve, percentile bootstrap over questions resampling both scores together	0.111	0.001	[0.0484, 0.177]	3998	5
The same holds under the labelled steps: the unspoken probability still beats the spoken step.	difference of two areas under the ROC curve, percentile bootstrap over questions resampling both scores together	0.108	0.001	[0.0506, 0.169]	3984	5
Under the forced choice the gap to the model’s own probability closes: the stated value is no longer clearly beaten by it.	difference of two areas under the ROC curve, percentile bootstrap over questions resampling both scores together	0.0496	0.19	[-0.0282, 0.124]	2000	5

Compute. Recorded GPU time for executed experiments is 0.1655 GPU-hours on one consumer GPU on the local machine, model served locally.

Code and data availability. A public code repository URL is not recorded in `project.yaml` ([links.github](https://github.com)). The per-example data of record that backs every reported number is provided under `results/real/` in the project repository: `figure_1_b4623192.csv`, `figure_2_4aacd28a.csv`, `questions.jsonl`, `run_1_b4623192_curve.csv`, `run_1_b4623192_results.csv`, `run_2_4aacd28a_curve.csv`, `run_2_4aacd28a_results.csv`, `experiments.json`.

References

- Groot, T., and Matias Valdenegro-Toro. 2024. *Overconfidence Is Key: Verbalized Uncertainty Evaluation in Large Language and Vision-Language Models*. <https://doi.org/10.48550/arXiv.2405.02917>.
- Kadavath, Saurav, Tom Conerly, Amanda Askell, et al. 2022. *Language Models (Mostly) Know What They Know*. <https://arxiv.org/abs/2207.05221>.
- Li, Chen, Xiaolin Hu, Songzhu Zheng, Jiawei Zhou, and Chao-Ran Chen. 2026. *ORCE: Order-Aware Alignment of Verbalized Confidence in Large Language Models*. <https://doi.org/10.48550/arXiv.2605.12446>.
- Lin, Stephanie, Jacob Hilton, and Owain Evans. 2022. *Teaching Models to Express Their Uncertainty in Words*. <https://arxiv.org/abs/2205.14334>.

- Tian, Katherine, E. Mitchell, Allan Zhou, et al. 2023. *Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback*. <https://doi.org/10.48550/arXiv.2305.14975>.
- Xiong, Miao, Zhiyuan Hu, Xinyang Lu, et al. 2024. *Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs*. <https://arxiv.org/abs/2306.13063>.