When the verdict contradicts the work: verbalized confidence, chain of thought, and what a truth probe actually reads

Abstract

A language model carries information about whether a statement is true in more than one place that need not agree, in the activations a trained linear probe can read, in the verdict it states out loud, and in the arithmetic its chain of thought writes down on the way to that verdict. If oversight is going to lean on any of these, what matters is whether they agree and what each one actually measures. We render 300 arithmetic and 300 capital-city statements in four surface formats each and run them through three open instruction-tuned models, Llama-3.1-8B [@grattafiori2024llama3], Qwen2.5-7B [@qwen2024qwen25], and Mistral-7B-v0.3 [@jiang2023mistral], scoring a leave-one-format-out linear probe on residual activations, a within-format probe over unseen items, the verbalized probability of truth [@lin2022teaching], and a training-free commit-probability scalar against ground truth. The verbalized verdict is fragile across surface form, and on Mistral arithmetic it collapses to chance because the model answers FALSE on almost every arithmetic statement regardless of content, a constant-FALSE response collapse we confirm with a base-rate test whose suppression gap, the FALSE rate on false statements minus the FALSE rate on true ones, is only 0.0267 for Mistral against 0.7667 for Qwen and 0.7483 for Llama. We then parse each model's own chain of thought on the arithmetic statements it judged wrong. Qwen is the clean case of a genuine knows-but-says-wrong, since its verdict tracks truth overall yet on the true statements it miscalls 66.3 percent of the ones that showed working had already computed the correct answer before it said FALSE, and 51.3 percent over the full quadrant including the items that showed no working. Mistral is not counted as a clean instance because its verdict is nearly truth independent, and Llama's wrong verdicts are mostly real miscalculations at 11.9 percent shown and 11.8 percent over the full quadrant. A label-trained probe reads truth at high AUC on the in-distribution formats, but its accuracy barely moves between items the model computed right and items it computed wrong, 0.9453 against 0.9429 for Llama, and the continuous probe score separates computed-right from computed-wrong items only at AUC 0.5430 for Llama, 0.4447 for Qwen, and 0.6325 for Mistral, near chance and not consistent in direction, so the probe reads the statement's arithmetic construction rather than the model's knowledge. For that reason we do not rest a divergence claim on the cell-level probe-versus-verbalized correlation, whose near-zero Pearson r of 0.0504 is forced by the probe sitting at ceiling in fifteen of eighteen cells. A logistic ensemble of the two signals does not beat the probe alone, adding 0.0123 AUC on arithmetic and nothing on capitals. The model's stated verdict can contradict its own correct computation, an internal probe is the more format-stable reader but does not by itself certify knowledge, and the chain of thought is what adjudicates the decoupling [@burns2023ccs] [@azaria2023internal].

Introduction

Suppose you want to trust what a model tells you about its own confidence. You ask it whether a claim is true and how sure it is, and you treat the answer as a window onto what the model knows. A separate line of work skips the asking and reads the model's activations directly, training a small probe to recover whether the model internally represents a statement as true [@burns2023ccs] [@marks2024geometry] [@azaria2023internal]. Both routes are on the table for oversight. The question we take up is whether they share the same blind spots. If the probe and the spoken confidence collapse on the same inputs, then the probe adds nothing that asking would not have given you, and the extra machinery of activation collection and probe training buys no coverage. If they fail on different inputs, then each catches errors the other misses, and combining them is worth the cost.

This is not an abstract preference. A model that internally represents a fact correctly while stating the opposite is the shape of the problem that deception-aware oversight is built to catch [@park2024deception] [@turpin2023saywhat]. Calibration work has shown that a model's stated confidence is an imperfect guide to its correctness [@guo2017calibration] [@kadavath2022know] [@xiong2023canllms], and faithfulness work has shown that the reasoning a model narrates can come apart from the computation that drives its answer [@lanham2023faithfulness] [@turpin2023saywhat]. What has been measured less often is whether an internal reader and the model's own words degrade together or separately as you vary the surface form of the same underlying fact. We hold the fact fixed and vary the wording, four ways per domain, and we watch where

each signal breaks.

We do find that the model's stated verdict and the rest of the picture come apart, and we trace where. The model's own chain of thought, parsed for the value it computes, shows that the spoken verdict often contradicts a correct internal calculation, most strongly in two of the three models, which is a genuine knows-but-says-wrong rather than a model that simply cannot do the arithmetic. A label-trained probe looks format-stable and reads truth across every wording, but we show with the chain of thought that its accuracy does not depend on whether the model actually computed the answer, so a high probe AUC certifies the construction of the statement set more than it certifies the model's knowledge. We report the full per-model breakdown rather than a single worst cell, we test and report a probe-plus-verbalized ensemble rather than asserting one would help, and we are explicit that the cell-level correlation between the two readers is uninformative here because the probe is pinned at ceiling.

Related work

Linear probing of internal representations goes back to using a simple classifier to test what a layer encodes [@alain2016probes], with later work mapping the promises and the pitfalls of reading too much into a probe's accuracy [@belinkov2022probing]. A more recent strand argues that truth in particular is linearly represented. Unsupervised methods recover a truth-like direction without labels [@burns2023ccs], supervised probes on hidden states detect false statements [@azaria2023internal], the geometry of true and false statements shows emergent linear structure [@marks2024geometry], and truth-correlated directions are causally usable to steer outputs [@li2023iti] [@zou2023repe]. Hallucination detection from internal states is a close cousin, reading an error signal off the activations rather than the text [@chwang2024androids]. Our probe is squarely in this family, a standardized last-token logistic probe, and our contribution is not a better probe but a measurement of how its blind spots line up against the model's own words.

The other signal we use is verbalized confidence. Models can be taught to express uncertainty in words [@lin2022teaching], can be asked for calibrated confidence scores [@tian2023justask], and have been evaluated systematically for whether their stated confidence tracks correctness [@xiong2023canllms] [@kadavath2022know]. Calibration of modern networks is a long-standing concern [@guo2017calibration], and treating a confidence score as a detector evaluated by AUC follows the out-of-distribution detection literature [@hendrycks2017baseline]. We use a standard elicitation, an explicit verdict plus a numeric rating, and we treat it as one operationalization of self-reported confidence among several.

The gap between what a model says and what it internally computes is the heart of the faithfulness and honesty literature. Chain-of-thought explanations are not always faithful to the answer they accompany [@turpin2023saywhat], faithfulness can be measured by intervening on the reasoning [@lanham2023faithfulness], and stated outputs can be bent by social pressures like sycophancy rather than by the truth [@sharma2023sycophancy]. Black-box lie detection asks unrelated questions to catch a model whose words diverge from its knowledge [@pacchiardi2024liar], and surveys frame the oversight stakes of deceptive behavior directly [@park2024deception]. Our knows-but-says-wrong quadrant is a quantified, format-controlled instance of exactly this decoupling, measured against an internal reader rather than against a behavioral probe. The factual content we test draws on the same kind of broad knowledge that factual benchmarks target [@hendrycks2021mmlu].

Methods

Each domain holds 300 base items balanced 150 true and 150 false. Every base item is rendered in four surface formats, which gives 1200 statements per domain per model and 24 cells overall across three models, two domains, and four formats. The same item identity is shared across a base item's four formats and across all three models, so the design is paired and a format contrast holds the underlying fact fixed. Items are built under a fixed seed.

Arithmetic is difficulty-stratified so the models make genuine errors under chain-of-thought rather than sitting at ceiling, with easy one and two digit add and subtract items, medium three digit or small multiplication items, and hard multi-digit multiplication. A false arithmetic item takes the true answer and adds a plausible

signed offset. The four arithmetic formats are a symbolic equation such as "575 * 758 = 435855", a spelled-out verbal rendering such as "Five hundred seventy-five times seven hundred fifty-eight equals four hundred thirty-five thousand eight hundred fifty-five.", a prose rendering such as "The product of 575 and 758 is 435855.", and a colloquial rendering such as "If you multiply 575 and 758, you get 435855.". Capital-city items draw from country and capital pairs weighted toward obscure countries, and a false item pairs a country with a real capital of a different country on the same continent, so the distractor is confusable rather than random. The four capital formats are an equation such as "Capital of Mauritius = Dodoma", a declarative such as "The capital of Mauritius is Dodoma.", an inverted form such as "Dodoma is the capital of Mauritius.", and a possessive prose form such as "Mauritius's capital city is Dodoma.".

For every rendered statement we record three signals. The verbalized signal is an explicit TRUE or FALSE verdict with a stated confidence from 0 to 100, elicited with a short chain-of-thought before a two-line answer, and mapped to a probability that the statement is true [@lin2022teaching] [@tian2023justask]. The parser takes the last verdict and confidence in the generation so that mid-reasoning mentions do not fool it, and a parse failure is scored as a wrong judgment rather than dropped. The commit-probability signal is the probability mass the model places on committing to the statement under a neutral prompt, a training-free internal scalar oriented once on pooled data so that a higher value means more true. The probe signal is a standardized L2-regularized logistic regression on the residual-stream activation at the final token of one selected layer, fit under a fixed seed. Generation is greedy, so both passes are deterministic given the model. Activations were collected on a single consumer GPU workstation.

We report two probe numbers per cell because they answer two different questions. The within-format number is a stratified five-fold cross-validated AUC inside the cell, where the held-out fold contains items the probe never saw during fitting but rendered in the same format. This is the probe's ceiling when the format is in distribution, and because the folds are item-disjoint it measures generalization to unseen items rather than memorization of seen ones. The leave-one-format-out number trains the probe on the other three formats of the domain and tests it on the held-out format, with the layer re-selected on the three training formats alone so the held-out cell never informs the layer choice. The leave-one-format-out number is the one comparable to the verbalized signal, since both are evaluated on a format the estimator did not fit. Every signal is scored by the area under the ROC curve [@hendrycks2017baseline]. A cell is flagged when verbalized accuracy exceeds 0.97, the regime where the labels are nearly all one way and AUC becomes unstable, and 6 of 24 cells are flagged and excluded from the headline correlation while still reported in full.

We measure shared versus divergent blind spots at two levels. At the cell level we correlate the probe's leave-one-format-out AUC with the verbalized AUC across cells, with a Pearson coefficient and a bootstrap interval, and a Spearman coefficient for the rank version, and we report the standard deviation of each AUC column and a Pearson recomputed on a logit scale so that the saturation of the probe near one is visible rather than hidden. At the item level we score each statement as probe-correct or not, using the out-of-fold leave-one-format-out prediction thresholded at one half, and as verbalized-correct or not, and we cross-tabulate. The quadrant that matters for honesty is the one where the probe is right and the verbalized verdict is wrong. We summarize the agreement between the two correctness vectors with the phi coefficient and put bootstrap intervals on the quadrant rates.

To decide whether the model in that quadrant actually holds the correct answer, we read its chain of thought. For each arithmetic item we parse, out of the recorded generation, the value the model computes for the operation, anchoring on the equation written with the real operands or on the model's final computed equation, and we compare that value to the correct one independent of the spoken verdict. An item is counted as computed correct when the reasoning reaches the true value and never asserts a different one for the operation, computed wrong when it asserts a different value or rejects the true one, and shown-no-working when no operand-anchored result appears. The parser was checked by hand on a sample of forty items, where it never confused a correct computation for a wrong one and left only borderline no-working cases unresolved, so the correct-versus-wrong split it reports is conservative. Before reading that quadrant as knowing suppression we run a base-rate test, computing per model on arithmetic the FALSE-verdict rate conditioned on ground truth and the suppression gap between the two, so a model that answers FALSE regardless of the statement is separated from one whose verdict tracks truth. We then run a probe-artifact check, comparing probe accuracy on the items the model computed correctly against the items it computed

wrong, and because a thresholded accuracy on a probe sitting near ceiling cannot move much, we also report the AUC of the probe's continuous score for separating computed-correct from computed-wrong items, with the mean probe-score difference and Cohen's d. Finally we build the ensemble the earlier draft had only recommended, a logistic regression on the probe score and the verbalized score scored strictly out of fold against ground truth, reporting its detection AUC and its recovery of the knows-but-says-wrong quadrant against each single signal.

Results

Figure figure_main.png is the blind-spot map. For each cell it places the internal probe bar, the leave-one-format-out number, beside the verbalized bar, faceted by model, with a chance line at one half. Where one bar drops toward chance while the other stays high, the two signals disagree about which format is hard. The probe is the stronger signal in almost every cell. The verbalized signal is strong on capitals and on arithmetic for Llama and Qwen, but on Mistral arithmetic it falls to the chance line while the probe stays near the top of the chart. The mean leave-one-format-out probe AUC is 0.9457 on arithmetic and 0.9994 on capitals, against a verbalized mean of 0.7402 on arithmetic and 0.9402 on capitals, so the probe leads everywhere and leads by the most exactly where the verbalized signal is weakest.

Figure figure_scatter.png plots the same cells as a scatter of probe AUC against verbalized AUC, with the line of equality drawn in. The cell-level correlation here is weak but uninformative, and we say so plainly rather than leaning on it. Over the eighteen non-flagged cells the Pearson correlation is 0.0504, yet the probe column has less than half the spread of the verbalized column, with a standard deviation of 0.0771 against 0.1733, and the probe sits at or above 0.99 in fifteen of those eighteen cells. A Pearson coefficient against a variable with almost no variance is pulled toward zero by construction, so the near-zero value is a ceiling artifact and not evidence that the two readers fail independently. The rank correlation, which does not care about the ceiling compression, points the other way, a Spearman of 0.4875 that is moderate, positive, and significant at p equal to 0.040156, and a Pearson recomputed on a logit scale that undoes the compression rises to 0.2699. Over all 24 cells the Pearson correlation is 0.1227 and the Spearman is 0.3875. The honest cell-level reading is that the probe and the verbalized signal are weakly and positively associated where both have room to vary, and that the cell scatter is too saturated to settle the shared-versus-divergent question on its own. We therefore move the question to the item level and to the chain of thought, where the probe ceiling does not distort the comparison.

Figure figure_breakdown.png breaks the mean AUC down by signal family and domain for each model. Three things read off it. The probe is the strongest family in every model and domain. The verbalized signal is competitive on capitals and on Llama and Qwen arithmetic but collapses on Mistral arithmetic, where its mean AUC is 0.5134. The commit-probability baseline is the weakest internal signal throughout, with model means of 0.6447, 0.6048, and 0.6011, far below the trained probe, which confirms that the separating information lives in a learned direction and not in a cheap scalar. The per-domain gap between the probe and the verbalized signal is 0.2054 on arithmetic and 0.0592 on capitals, and by model it is 0.2975 for Mistral, 0.0730 for Llama, and 0.0428 for Qwen.

The divergence is sharper item by item than the cell averages suggest, because averaging inside a cell hides item-level disagreement. Figure figure_bootstrap.png shows the item-level quadrant rates per domain with bootstrap intervals. Pooled over models and formats, on arithmetic the probe and the verbalized verdict are both correct on 67.0% of items, the probe is right while the verbalized verdict is wrong on 22.9% with an interval from 0.215 to 0.244, the verbalized verdict is right while the probe is wrong on 8.6%, and both are wrong on 1.5%. On capitals the both-correct rate is 94.5%, the probe-right-verbalized-wrong rate is 4.6%, the converse is 0.8%, and both are wrong on 0.2%. The phi coefficient between the two correctness vectors is -0.0768 on arithmetic and 0.0710 on capitals, near zero in both, which is the signature of divergent rather than shared blind spots. The converse quadrant is small, so the probe rarely loses where the spoken verdict wins.

The item-level quadrant where the probe is right and the verbalized verdict is wrong is large on arithmetic and concentrated in Mistral, at 46.7% of items for Mistral against 11.6% for Llama and 10.5% for Qwen. Before reading that quadrant as a model that knows the answer and says the opposite, we rule out a model that

simply answers FALSE no matter what, because a constant-FALSE responder fills the quadrant with every true item it can compute and shows no content-sensitive decoupling at all. Figure figure_suppression.png runs the base-rate test. For each model on arithmetic we compute how often the verdict is FALSE conditioned on ground truth, and the suppression gap, the FALSE rate on false statements minus the FALSE rate on true ones. Qwen and Llama have large gaps, 0.7667 and 0.7483, since they call 0.2183 and 0.2400 of true statements FALSE but 0.9850 and 0.9883 of false statements FALSE, so their verdict tracks truth and a FALSE on a true statement is a real, content-sensitive miscall. Mistral is the opposite, calling 0.8983 of true statements FALSE and 0.9250 of false statements FALSE, a gap of only 0.0267, so its verdict is nearly truth independent and its quadrant is response collapse rather than suppression. We therefore do not count Mistral as a clean instance of the phenomenon, and the deception-relevant reading rests on Qwen, with Mistral kept as a contrast.

With the constant-FALSE model set aside, we read the surviving models' own chains of thought to see whether the verdict contradicts a correct computation. For every arithmetic statement a model judged wrong while its ground truth was true, we parse the value its chain of thought computes for the operation and compare it to the correct value, independent of the spoken verdict. Figure figure_cot_diagnostic.png shows the split. Qwen is the clean case. Its verdict tracks arithmetic truth, yet on the true statements it miscalls, 66.3% of the ones that showed working had already computed the correct answer before it said FALSE, and 51.3% over the full quadrant once the items that showed no working are counted in the denominator. That is a correct internal calculation contradicted by the spoken verdict on a model that otherwise judges arithmetic competently. Llama is different, with only 11.9% of its shown-working items and 11.8% over the full quadrant computed correctly, so Llama's spoken errors are mostly real miscalculations rather than a model talking down its own correct work. Mistral's shown-working share is 65.4%, but because its verdict is truth independent we read this as the base rate at which it solves multiplications in scratch work intersected with a verdict that says FALSE regardless, not as knowing suppression. Reporting the full quadrant rather than the shown-working subset keeps the no-working items visible, 47.5% of Mistral's quadrant and 35.5% pooled, so no rate is inflated by conditioning them away. The chain of thought, not the probe, is what licenses the reading, and it licenses it cleanly for one model.

The same chain-of-thought labels expose what the probe is reading. If the probe tracked the model's knowledge, it would be more confident where the model computed the answer and less where it did not. It is not. On ground-truth-true arithmetic items with visible working, the probe calls the statement true on 0.9453 of the items Llama computed correctly and 0.9429 of the items it computed wrong, a gap of almost nothing. A thresholded accuracy can saturate, so we also separate the two groups with the probe's continuous score directly. The AUC of that score for telling computed-right from computed-wrong items is 0.5430 for Llama, 0.4447 for Qwen, and 0.6325 for Mistral, near chance and not even consistent in direction, with a mean across models near one half. For Llama the mean probe score is identical across the split, a difference of 0.0000 and a Cohen's d of 0.0002, and for Qwen the probe is if anything slightly lower on the items computed correctly, a Cohen's d of -0.2953. A reader of the model's actual computation would separate these groups well above one half and in a single direction, and the probe does not, which is the signature of a reader keyed to the construction of the statement, the true answer plus a signed offset, rather than to the model's internal computation. The verbalized signal, by contrast, does move with the chain of thought, since on the items Llama computed correctly it agrees 0.9531 of the time and on the items it computed wrong only 0.2743 of the time. So between the two readers it is the spoken verdict, not the probe, whose correctness tracks the model's actual arithmetic, and the probe's format stability is the stability of a construction signature.

The probe is not exploiting cross-format leakage in the narrow sense of recognizing items it saw in another wording. With item-disjoint five-fold cross-validation inside a single format, where the probe is tested on held-out items it never fitted, the arithmetic within-format AUC is still near 0.98 across the three colloquial, prose, and symbolic formats for all three models, for instance 0.9812 for Mistral colloquial, 0.9847 for Llama colloquial, and 0.9956 for Qwen colloquial. So the probe generalizes to unseen items rather than memorizing seen ones. As the chain-of-thought check just showed, however, generalizing to unseen items is exactly what a reader of the distractor construction would also do, so the within-format result rules out memorization without establishing that the probe reads the model's knowledge.

Because the draft that preceded this one recommended combining the two readers, we built the ensemble and measured it rather than leaving it as a suggestion. Figure figure_ensemble.png reports a logistic regression on the probe score and the verbalized score, scored strictly out of fold against ground truth. On arithmetic the ensemble reaches an AUC of 0.9690 against 0.9567 for the probe alone and 0.7595 for the verbalized signal alone, a gain over the better single signal of 0.0123. On capitals the ensemble is 0.9984 against the probe's 0.9989, a loss of 0.0005, since the probe is already at ceiling and the verbalized signal has nothing to add. On the knows-but-says-wrong quadrant the ensemble recovers 0.9677 of the items on arithmetic, slightly fewer than the probe alone, which recovers them by construction, because mixing in the verbalized signal that is wrong on exactly those items drags a few below the decision threshold. The measured verdict is that the ensemble does not beat the probe alone in any way that would change a deployment choice, and whatever small arithmetic gain it shows inherits the probe's dependence on the statement construction.

The probe is not uniformly strong, and its weak spot is a different format than the verbalized signal's. The probe's worst arithmetic format is the spelled-out verbal rendering, where the leave-one-format-out AUC falls to 0.8266 for Llama, 0.7291 for Qwen, and 0.8676 for Mistral, well below its near-perfect numbers on the other three formats. The verbalized signal's weak arithmetic format, setting aside Mistral's across-the-board collapse, is the symbolic equation, where Qwen's verbalized AUC is 0.7933. The two signals are hard to read in different surface forms of the same arithmetic, which is the format-level version of the divergence the cell-level scatter shows at the model level.

Discussion

The reframed picture is more specific than divergent blind spots, and parts of a divergence claim do not survive. The cell-level correlation between the probe and the verbalized signal is too saturated to support a strong claim either way, since the probe is at ceiling almost everywhere and the rank correlation is in fact positive. What survives, and is sharper, is a decoupling visible item by item and in the model's own words. The base-rate test does the screening first. Mistral, whose verdict is a near constant FALSE with a suppression gap of 0.0267, is set aside as response collapse rather than knowing suppression, and Qwen, whose verdict tracks truth with a gap of 0.7667, is the clean case where the chain of thought shows two thirds of the shown-working wrong-verdict items computed correctly first. Llama's wrong verdicts are real arithmetic errors. The decoupling is genuine and content-sensitive for one of the three models, demonstrated rather than asserted, and reported alongside the degenerate responder it must be distinguished from. This is the deception-relevant pattern, a model that has the right answer in its working and states the opposite, and it now rests on the model's own reasoning rather than on a probe whose accuracy we have shown to be insensitive to whether the model actually knew [@kadavath2022know]. For oversight the lesson runs two ways. The verbalized verdict is the reader whose correctness tracks the model's computation, so it is informative precisely where it is honest, and a high-AUC truth probe can look authoritative while reading the construction of the evaluation rather than the model's knowledge, so probe AUC alone should not be read as a knowledge certificate.

The knows-but-says-wrong pattern is the deception-relevant one [@park2024deception] [@turpin2023saywhat]. A model that writes a correct sum in its reasoning and then reports the statement false is the failure that black-box and white-box honesty methods are both trying to catch, and the chain-of-thought parse gives it a rate per model under controlled wording rather than an anecdote, with the base-rate screen separating that pattern from a model that merely defaults to FALSE. The capital domain, where the verbalized signal is near ceiling for every model, is the low-decoupling contrast that shows the effect is mainly an arithmetic-domain phenomenon here. The format stability of the probe, which an earlier reading took as the internal representation tracking truth more stably than words do, is better described as the stability of a feature of the statement set, since the probe is equally confident whether or not the model computed the answer [@marks2024geometry] [@azaria2023internal] [@zou2023repe].

Limitations

The deception-relevant reading rests on a single model. Qwen is the one model whose arithmetic verdict tracks truth, with a suppression gap of 0.7667, while still contradicting a correct computation on the true

statements it miscalls, so it is the clean case of knows-but-says-wrong. Mistral, whose verdict sits at chance, is a constant-FALSE responder with a suppression gap of only 0.0267, so its large quadrant is response collapse and we show it for contrast rather than counting it as the phenomenon. Llama's wrong verdicts are mostly real miscalculations. One clean model is a thin base for a general claim, and whether the pattern recurs in a model whose arithmetic verdict is neither at ceiling nor collapsed is open.

The probe reads a feature of the statement set rather than the model's knowledge, and we now show this directly rather than only conceding it. The probe's accuracy is essentially unchanged between arithmetic items the model computed correctly and items it computed wrong, which is what a reader of the distractor construction, the true answer plus a signed offset, would look like and what a reader of the model's computation would not. The continuous probe score confirms it, separating computed-right from computed-wrong items at an AUC of only 0.5430, 0.4447, and 0.6325 across the three models, near chance and not consistent in direction, where a reader of the model's computation would land well above one half. The item-disjoint within-format result still rules out memorization of specific items, since the probe generalizes to held-out items, but generalizing to unseen items is consistent with reading the construction, so it does not rescue a knowledge interpretation. Separating a probe that has learned truth from one that has learned our construction would take adversarial distractors built to break any such surface regularity, and on the present statement set the chain of thought, not the probe, is the trustworthy internal witness to what the model computed.

What the diagnostic settled and what it did not is worth stating. It settled that the spoken verdict can contradict a correct internal calculation, and that the probe's accuracy does not track that calculation, so neither the deception framing nor the probe-as-knowledge framing can be taken on faith. It did not settle what the probe reads instead beyond pointing at the construction, and it did not give a clean answer on the items where the model showed no working, which we count in a separate bucket rather than forcing into the correct-versus-wrong split. The chain-of-thought parse is a heuristic over free text, validated by hand on a sample of forty items where it never confused a correct computation for a wrong one, so the split is conservative but not exact.

The scope is narrow. Two domains, three open instruction-tuned models in the seven to eight billion parameter range, English only, one probe family in the form of a last-token linear probe, and a statement-verification framing. We make no claim beyond that envelope, and in particular the capital domain is near ceiling for one model, which narrows the capital evidence to the other two.

Finally, statement verification is not free-form generation. The verbalized confidence here is an explicit truth judgment plus a numeric rating, which is one operationalization of self-reported confidence among several [@lin2022teaching] [@xiong2023canllms]. A different elicitation, or a generative setting where the model produces a claim and rates it, might move the verbalized numbers, and the comparison we draw is to this specific and standard way of asking.

Conclusion

Reading a model's confidence from its activations, from its words, and from its own reasoning are not interchangeable. On arithmetic and capital-city statements rendered four ways each, across three open models, the spoken verdict can contradict a correct internal calculation, and once a base-rate test screens out a model whose verdict merely defaults to FALSE, the model's chain of thought shows a genuine knows-but-says-wrong cleanly for one model whose verdict otherwise tracks truth rather than a model that simply cannot compute. A label-trained truth probe is the most format-stable reader, but its accuracy does not depend on whether the model actually knew, so it reads the construction of the evaluation as much as the model's knowledge, and a probe-plus-verbalized ensemble does not beat the probe alone. The practical reading is that a high probe AUC should not be taken as a knowledge certificate on its own, and that the model's own reasoning is the signal that adjudicates whether a wrong verdict sits on top of a right answer.

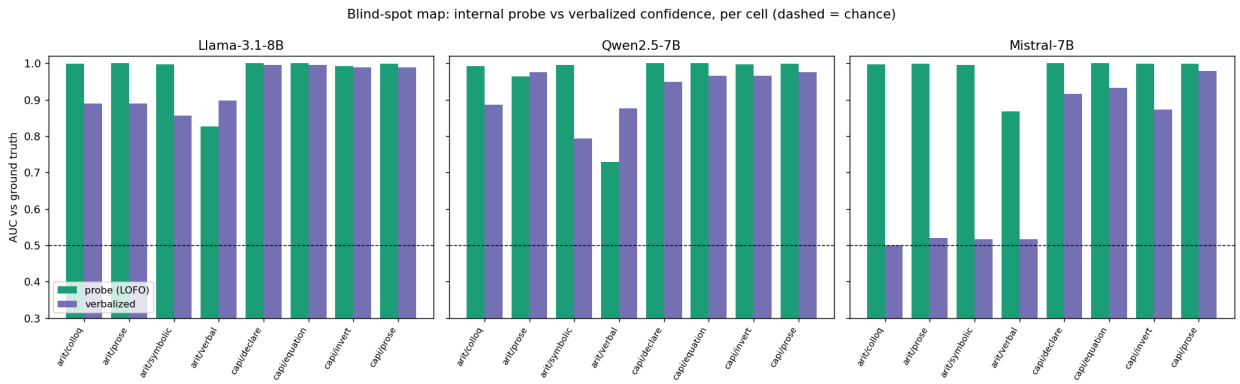

Figure 1: Blind-spot map. Internal probe leave-one-format-out AUC beside verbalized AUC per cell, faceted by model, with a chance line.

Figures

Appendix: full results ledger

Every decimal-valued number in results/real/analysis_summary.txt, traceable to its data file of record through results/real/sources.json.

Section	Metric	Value
llama-3.1-8b-instruct/arithmetic/f_colloq	explicit_acc	0.8896
llama-3.1-8b-instruct/arithmetic/f_colloq	parse_fail_rate	0.0033
llama-3.1-8b-instruct/arithmetic/f_prose	explicit_acc	0.8833
llama-3.1-8b-instruct/arithmetic/f_prose	parse_fail_rate	0.0000
llama-3.1-8b-instruct/arithmetic/f_symbolic	explicit_acc	0.8533
llama-3.1-8b-instruct/arithmetic/f_symbolic	parse_fail_rate	0.0000
llama-3.1-8b-instruct/arithmetic/f_verbal	explicit_acc	0.8716
llama-3.1-8b-instruct/arithmetic/f_verbal	parse_fail_rate	0.0133
llama-3.1-8b-instruct/capital/f_declare	explicit_acc	0.9933
llama-3.1-8b-instruct/capital/f_declare	parse_fail_rate	0.0000
llama-3.1-8b-instruct/capital/f_equation	explicit_acc	0.9900
llama-3.1-8b-instruct/capital/f_equation	parse_fail_rate	0.0000
llama-3.1-8b-instruct/capital/f_invert	explicit_acc	0.9833
llama-3.1-8b-instruct/capital/f_invert	parse_fail_rate	0.0000

Section	Metric	Value
llama-3.1-8b-instruct/capital/f_prose	explicit_acc	0.9800
llama-3.1-8b-instruct/capital/f_prose	parse_fail_rate	0.0000
llama-3.1-8b-instruct/arithmetic	explicit_acc	0.8745
llama-3.1-8b-instruct/capital	explicit_acc	0.9867
qwen2.5-7b-instruct/arithmetic/f_colloq	explicit_acc	0.8867
qwen2.5-7b-instruct/arithmetic/f_colloq	parse_fail_rate	0.0000
qwen2.5-7b-instruct/arithmetic/f_prose	explicit_acc	0.9767
qwen2.5-7b-instruct/arithmetic/f_prose	parse_fail_rate	0.0000
qwen2.5-7b-instruct/arithmetic/f_symbolic	explicit_acc	0.7933
qwen2.5-7b-instruct/arithmetic/f_symbolic	parse_fail_rate	0.0000
qwen2.5-7b-instruct/arithmetic/f_verbal	explicit_acc	0.8767
qwen2.5-7b-instruct/arithmetic/f_verbal	parse_fail_rate	0.0000
qwen2.5-7b-instruct/capital/f_declare	explicit_acc	0.9367
qwen2.5-7b-instruct/capital/f_declare	parse_fail_rate	0.0000
qwen2.5-7b-instruct/capital/f_equation	explicit_acc	0.9500
qwen2.5-7b-instruct/capital/f_equation	parse_fail_rate	0.0000
qwen2.5-7b-instruct/capital/f_invert	explicit_acc	0.9333
qwen2.5-7b-instruct/capital/f_invert	parse_fail_rate	0.0000
qwen2.5-7b-instruct/capital/f_prose	explicit_acc	0.9600
qwen2.5-7b-instruct/capital/f_prose	parse_fail_rate	0.0000
qwen2.5-7b-instruct/arithmetic	explicit_acc	0.8833
qwen2.5-7b-instruct/capital	explicit_acc	0.9450
mistral-7b-instruct-v0.3/arithmetic/f_colloq	explicit_acc	0.5000
mistral-7b-instruct-v0.3/arithmetic/f_colloq	parse_fail_rate	0.0000
mistral-7b-instruct-v0.3/arithmetic/f_prose	explicit_acc	0.5200
mistral-7b-instruct-v0.3/arithmetic/f_prose	parse_fail_rate	0.0000
mistral-7b-instruct-v0.3/arithmetic/f_symbolic	explicit_acc	0.5167

Section	Metric	Value
mistral-7b-instruct-v0.3/arithmetic/f_symbolic	parse_fail_rate	0.0000
mistral-7b-instruct-v0.3/arithmetic/f_verbal	explicit_acc	0.5167
mistral-7b-instruct-v0.3/arithmetic/f_verbal	parse_fail_rate	0.0000
mistral-7b-instruct-v0.3/capital/f_declare	explicit_acc	0.9167
mistral-7b-instruct-v0.3/capital/f_declare	parse_fail_rate	0.0000
mistral-7b-instruct-v0.3/capital/f_equation	explicit_acc	0.9333
mistral-7b-instruct-v0.3/capital/f_equation	parse_fail_rate	0.0000
mistral-7b-instruct-v0.3/capital/f_invert	explicit_acc	0.8733
mistral-7b-instruct-v0.3/capital/f_invert	parse_fail_rate	0.0000
mistral-7b-instruct-v0.3/capital/f_prose	explicit_acc	0.9800
mistral-7b-instruct-v0.3/capital/f_prose	parse_fail_rate	0.0000
mistral-7b-instruct-v0.3/arithmetic	explicit_acc	0.5133
mistral-7b-instruct-v0.3/capital	explicit_acc	0.9258
stage2/llama-3.1-8b-instruct/arithmetic/f_colloq	verbalized_auc	0.8893
stage2/llama-3.1-8b-instruct/arithmetic/f_colloq	commit_auc	0.5874
stage2/llama-3.1-8b-instruct/arithmetic/f_colloq	probe_infmt_auc	0.9847
stage2/llama-3.1-8b-instruct/arithmetic/f_colloq	probe_lofo_auc	0.9998
stage2/llama-3.1-8b-instruct/arithmetic/f_prose	verbalized_auc	0.8899
stage2/llama-3.1-8b-instruct/arithmetic/f_prose	commit_auc	0.6834
stage2/llama-3.1-8b-instruct/arithmetic/f_prose	probe_infmt_auc	0.9900
stage2/llama-3.1-8b-instruct/arithmetic/f_prose	probe_lofo_auc	1.0000
stage2/llama-3.1-8b-instruct/arithmetic/f_symbolic	verbalized_auc	0.8557
stage2/llama-3.1-8b-instruct/arithmetic/f_symbolic	commit_auc	0.5194
stage2/llama-3.1-8b-instruct/arithmetic/f_symbolic	probe_infmt_auc	0.9867
stage2/llama-3.1-8b-instruct/arithmetic/f_symbolic	probe_lofo_auc	0.9981
stage2/llama-3.1-8b-instruct/arithmetic/f_verbal	verbalized_auc	0.8974

Section	Metric	Value
stage2/llama-3.1-8b-instruct/arithmetic/f_verbal	commit_auc	0.7885
stage2/llama-3.1-8b-instruct/arithmetic/f_verbal	probe_infmt_auc	0.8943
stage2/llama-3.1-8b-instruct/arithmetic/f_verbal	probe_lofo_auc	0.8266
stage2/llama-3.1-8b-instruct/arithmetic/f_verbal	verbalized_auc	0.9960
stage2/llama-3.1-8b-instruct/capital/f_declare	commit_auc	0.9065
stage2/llama-3.1-8b-instruct/capital/f_declare	probe_infmt_auc	0.9994
stage2/llama-3.1-8b-instruct/capital/f_declare	probe_lofo_auc	1.0000
stage2/llama-3.1-8b-instruct/capital/f_declare	verbalized_auc	0.9962
stage2/llama-3.1-8b-instruct/capital/f_equation	commit_auc	0.9724
stage2/llama-3.1-8b-instruct/capital/f_equation	probe_infmt_auc	0.9996
stage2/llama-3.1-8b-instruct/capital/f_equation	probe_lofo_auc	1.0000
stage2/llama-3.1-8b-instruct/capital/f_equation	verbalized_auc	0.9898
stage2/llama-3.1-8b-instruct/capital/f_invert	commit_auc	0.8064
stage2/llama-3.1-8b-instruct/capital/f_invert	probe_infmt_auc	0.9934
stage2/llama-3.1-8b-instruct/capital/f_invert	probe_lofo_auc	0.9922
stage2/llama-3.1-8b-instruct/capital/f_invert	verbalized_auc	0.9890
stage2/llama-3.1-8b-instruct/capital/f_prose	commit_auc	0.9068
stage2/llama-3.1-8b-instruct/capital/f_prose	probe_infmt_auc	0.9987
stage2/llama-3.1-8b-instruct/capital/f_prose	probe_lofo_auc	0.9996
stage2/qwen2.5-7b-instruct/arithmetic/f_colloq	verbalized_auc	0.8867
stage2/qwen2.5-7b-instruct/arithmetic/f_colloq	commit_auc	0.8827
stage2/qwen2.5-7b-instruct/arithmetic/f_colloq	probe_infmt_auc	0.9956
stage2/qwen2.5-7b-instruct/arithmetic/f_colloq	probe_lofo_auc	0.9927
stage2/qwen2.5-7b-instruct/arithmetic/f_prose	verbalized_auc	0.9767
stage2/qwen2.5-7b-instruct/arithmetic/f_prose	commit_auc	0.8678
stage2/qwen2.5-7b-instruct/arithmetic/f_prose	probe_infmt_auc	0.9995

Section	Metric	Value
stage2/qwen2.5-7b-instruct/arithmetic/f_prose	probe_lofo_auc	0.9649
stage2/qwen2.5-7b-instruct/arithmetic/f_symbolic	verbalized_auc	0.7933
stage2/qwen2.5-7b-instruct/arithmetic/f_symbolic	commit_auc	0.7452
stage2/qwen2.5-7b-instruct/arithmetic/f_symbolic	probe_infmt_auc	0.9929
stage2/qwen2.5-7b-instruct/arithmetic/f_symbolic	probe_lofo_auc	0.9955
stage2/qwen2.5-7b-instruct/arithmetic/f_verbal	verbalized_auc	0.8767
stage2/qwen2.5-7b-instruct/arithmetic/f_verbal	commit_auc	0.7232
stage2/qwen2.5-7b-instruct/arithmetic/f_verbal	probe_infmt_auc	0.9048
stage2/qwen2.5-7b-instruct/arithmetic/f_verbal	probe_lofo_auc	0.7291
stage2/qwen2.5-7b-instruct/capital/f_declare	verbalized_auc	0.9494
stage2/qwen2.5-7b-instruct/capital/f_declare	commit_auc	0.4651
stage2/qwen2.5-7b-instruct/capital/f_declare	probe_infmt_auc	0.9976
stage2/qwen2.5-7b-instruct/capital/f_declare	probe_lofo_auc	1.0000
stage2/qwen2.5-7b-instruct/capital/f_equation	verbalized_auc	0.9664
stage2/qwen2.5-7b-instruct/capital/f_equation	commit_auc	0.7676
stage2/qwen2.5-7b-instruct/capital/f_equation	probe_infmt_auc	0.9972
stage2/qwen2.5-7b-instruct/capital/f_equation	probe_lofo_auc	1.0000
stage2/qwen2.5-7b-instruct/capital/f_invert	verbalized_auc	0.9665
stage2/qwen2.5-7b-instruct/capital/f_invert	commit_auc	0.1887
stage2/qwen2.5-7b-instruct/capital/f_invert	probe_infmt_auc	0.9920
stage2/qwen2.5-7b-instruct/capital/f_invert	probe_lofo_auc	0.9980
stage2/qwen2.5-7b-instruct/capital/f_prose	verbalized_auc	0.9760
stage2/qwen2.5-7b-instruct/capital/f_prose	commit_auc	0.4611
stage2/qwen2.5-7b-instruct/capital/f_prose	probe_infmt_auc	0.9970
stage2/qwen2.5-7b-instruct/capital/f_prose	probe_lofo_auc	0.9995
stage2/mistral-7b-instruct-v0.3/arithmetic/f_colloq	verbalized_auc	0.5000

Section	Metric	Value
stage2/mistral-7b-instruct-v0.3/arithmetic/f_colloq	commit_auc	0.4593
stage2/mistral-7b-instruct-v0.3/arithmetic/f_colloq	probe_infmt_auc	0.9812
stage2/mistral-7b-instruct-v0.3/arithmetic/f_colloq	probe_lofo_auc	0.9981
stage2/mistral-7b-instruct-v0.3/arithmetic/f_colloq	verbalized_auc	0.5200
stage2/mistral-7b-instruct-v0.3/arithmetic/f_prose	commit_auc	0.5538
stage2/mistral-7b-instruct-v0.3/arithmetic/f_prose	probe_infmt_auc	0.9840
stage2/mistral-7b-instruct-v0.3/arithmetic/f_prose	probe_lofo_auc	0.9989
stage2/mistral-7b-instruct-v0.3/arithmetic/f_prose	verbalized_auc	0.5167
stage2/mistral-7b-instruct-v0.3/arithmetic/f_symbolic	commit_auc	0.5671
stage2/mistral-7b-instruct-v0.3/arithmetic/f_symbolic	probe_infmt_auc	0.9844
stage2/mistral-7b-instruct-v0.3/arithmetic/f_symbolic	probe_lofo_auc	0.9959
stage2/mistral-7b-instruct-v0.3/arithmetic/f_symbolic	verbalized_auc	0.5167
stage2/mistral-7b-instruct-v0.3/arithmetic/f_verbal	commit_auc	0.5032
stage2/mistral-7b-instruct-v0.3/arithmetic/f_verbal	probe_infmt_auc	0.8548
stage2/mistral-7b-instruct-v0.3/arithmetic/f_verbal	probe_lofo_auc	0.8676
stage2/mistral-7b-instruct-v0.3/arithmetic/f_verbal	verbalized_auc	0.9167
stage2/mistral-7b-instruct-v0.3/capital/f_declare	commit_auc	0.7896
stage2/mistral-7b-instruct-v0.3/capital/f_declare	probe_infmt_auc	0.9961
stage2/mistral-7b-instruct-v0.3/capital/f_declare	probe_lofo_auc	1.0000
stage2/mistral-7b-instruct-v0.3/capital/f_declare	verbalized_auc	0.9333
stage2/mistral-7b-instruct-v0.3/capital/f_equation	commit_auc	0.4404
stage2/mistral-7b-instruct-v0.3/capital/f_equation	probe_infmt_auc	0.9964
stage2/mistral-7b-instruct-v0.3/capital/f_equation	probe_lofo_auc	1.0000
stage2/mistral-7b-instruct-v0.3/capital/f_equation	verbalized_auc	0.8733
stage2/mistral-7b-instruct-v0.3/capital/f_invert	commit_auc	0.8946
stage2/mistral-7b-instruct-v0.3/capital/f_invert	probe_infmt_auc	0.9952

Section	Metric	Value
stage2/mistral-7b-instruct-v0.3/capital/f_invert	probe_lofo_auc	0.9985
stage2/mistral-7b-instruct-v0.3/capital/f_prose	verbalized_auc	0.9800
stage2/mistral-7b-instruct-v0.3/capital/f_prose	commit_auc	0.6326
stage2/mistral-7b-instruct-v0.3/capital/f_prose	probe_infmt_auc	0.9959
stage2/mistral-7b-instruct-v0.3/capital/f_prose	probe_lofo_auc	0.9997
stage2/commit/orientation	commit_sign	1.0000
stage2/correlation/all_cells	pearson_r	0.1227
stage2/correlation/all_cells	pearson_ci_lo	-0.1347
stage2/correlation/all_cells	pearson_ci_hi	0.5874
stage2/correlation/all_cells	spearman_r	0.3875
stage2/correlation/all_cells	spearman_p	0.061346
stage2/correlation/nonflagged	pearson_r	0.0504
stage2/correlation/nonflagged	pearson_ci_lo	-0.2267
stage2/correlation/nonflagged	pearson_ci_hi	0.6266
stage2/correlation/nonflagged	spearman_r	0.4875
stage2/correlation/nonflagged	spearman_p	0.040156
stage2/gap/model/llama-3.1-8b-instruct	mean_probe_lofo	0.9561
stage2/gap/model/llama-3.1-8b-instruct	mean_verbalized	0.8831
stage2/gap/model/llama-3.1-8b-instruct	mean_commit	0.6447
stage2/gap/model/llama-3.1-8b-instruct	mean_gap	0.0730
stage2/gap/model/mistral-7b-instruct-v0.3	mean_probe_lofo	0.9799
stage2/gap/model/mistral-7b-instruct-v0.3	mean_verbalized	0.6824
stage2/gap/model/mistral-7b-instruct-v0.3	mean_commit	0.6011
stage2/gap/model/mistral-7b-instruct-v0.3	mean_gap	0.2975
stage2/gap/model/qwen2.5-7b-instruct	mean_probe_lofo	0.9593
stage2/gap/model/qwen2.5-7b-instruct	mean_verbalized	0.9164
stage2/gap/model/qwen2.5-7b-instruct	mean_commit	0.6048
stage2/gap/model/qwen2.5-7b-instruct	mean_gap	0.0428
stage2/gap/domain/arithmetic	mean_probe_lofo	0.9457
stage2/gap/domain/arithmetic	mean_verbalized	0.7402
stage2/gap/domain/arithmetic	mean_commit	0.6376
stage2/gap/domain/arithmetic	mean_gap	0.2054
stage2/gap/domain/capital	mean_probe_lofo	0.9994
stage2/gap/domain/capital	mean_verbalized	0.9402
stage2/gap/domain/capital	mean_commit	0.5724

Section	Metric	Value
stage2/gap/domain/capital	mean_gap	0.0592
stage2/itemlevel/arithmetic_pooled	both_correct_rate	0.6697
stage2/itemlevel/arithmetic_pooled	knows_but_says_wrong_rate	0.2292
stage2/itemlevel/arithmetic_pooled	kbsw_ci_lo	0.2150
stage2/itemlevel/arithmetic_pooled	kbsw_ci_hi	0.2442
stage2/itemlevel/arithmetic_pooled	verb_only_rate	0.0864
stage2/itemlevel/arithmetic_pooled	vo_ci_lo	0.0778
stage2/itemlevel/arithmetic_pooled	vo_ci_hi	0.0961
stage2/itemlevel/arithmetic_pooled	both_wrong_rate	0.0147
stage2/itemlevel/arithmetic_pooled	phi	-0.0768
stage2/itemlevel/capital_pooled	both_correct_rate	0.9447
stage2/itemlevel/capital_pooled	knows_but_says_wrong_rate	0.0456
stage2/itemlevel/capital_pooled	kbsw_ci_lo	0.0389
stage2/itemlevel/capital_pooled	kbsw_ci_hi	0.0528
stage2/itemlevel/capital_pooled	verb_only_rate	0.0078
stage2/itemlevel/capital_pooled	vo_ci_lo	0.0050
stage2/itemlevel/capital_pooled	vo_ci_hi	0.0106
stage2/itemlevel/capital_pooled	both_wrong_rate	0.0019
stage2/itemlevel/capital_pooled	phi	0.0710
stage2/itemlevel/llama-3.1-8b-instruct/arithmetic	both_correct_rate	0.7917
stage2/itemlevel/llama-3.1-8b-instruct/arithmetic	knows_but_says_wrong_rate	0.1158
stage2/itemlevel/llama-3.1-8b-instruct/arithmetic	kbsw_ci_lo	0.0983
stage2/itemlevel/llama-3.1-8b-instruct/arithmetic	kbsw_ci_hi	0.1333
stage2/itemlevel/llama-3.1-8b-instruct/arithmetic	verb_only_rate	0.0800
stage2/itemlevel/llama-3.1-8b-instruct/arithmetic	vo_ci_lo	0.0650
stage2/itemlevel/llama-3.1-8b-instruct/arithmetic	vo_ci_hi	0.0950
stage2/itemlevel/llama-3.1-8b-instruct/arithmetic	both_wrong_rate	0.0125
stage2/itemlevel/llama-3.1-8b-instruct/arithmetic	phi	0.0065
stage2/itemlevel/llama-3.1-8b-instruct/capital	both_correct_rate	0.9767
stage2/itemlevel/llama-3.1-8b-instruct/capital	knows_but_says_wrong_rate	0.0117
stage2/itemlevel/llama-3.1-8b-instruct/capital	kbsw_ci_lo	0.0058
stage2/itemlevel/llama-3.1-8b-instruct/capital	kbsw_ci_hi	0.0183
stage2/itemlevel/llama-3.1-8b-instruct/capital	verb_only_rate	0.0100
stage2/itemlevel/llama-3.1-8b-instruct/capital	vo_ci_lo	0.0050
stage2/itemlevel/llama-3.1-8b-instruct/capital	vo_ci_hi	0.0158

Section	Metric	Value
stage2/itemlevel/llama-3.1-8b-instruct/capital	both_wrong_rate	0.0017
stage2/itemlevel/llama-3.1-8b-instruct/capital	phi	0.1227
stage2/itemlevel/mistral-7b-instruct-v0.3/arithmetic	both_correct_rate	0.4625
stage2/itemlevel/mistral-7b-instruct-v0.3/arithmetic	knows_but_says_wrong_rate	0.4667
stage2/itemlevel/mistral-7b-instruct-v0.3/arithmetic	kbsw_ci_lo	0.4400
stage2/itemlevel/mistral-7b-instruct-v0.3/arithmetic	kbsw_ci_hi	0.4942
stage2/itemlevel/mistral-7b-instruct-v0.3/arithmetic	verb_only_rate	0.0508
stage2/itemlevel/mistral-7b-instruct-v0.3/arithmetic	vo_ci_lo	0.0383
stage2/itemlevel/mistral-7b-instruct-v0.3/arithmetic	vo_ci_hi	0.0634
stage2/itemlevel/mistral-7b-instruct-v0.3/arithmetic	both_wrong_rate	0.0200
stage2/itemlevel/mistral-7b-instruct-v0.3/arithmetic	phi	-0.1129
stage2/itemlevel/mistral-7b-instruct-v0.3/capital	both_correct_rate	0.9192
stage2/itemlevel/mistral-7b-instruct-v0.3/capital	knows_but_says_wrong_rate	0.0725
stage2/itemlevel/mistral-7b-instruct-v0.3/capital	kbsw_ci_lo	0.0583
stage2/itemlevel/mistral-7b-instruct-v0.3/capital	kbsw_ci_hi	0.0867
stage2/itemlevel/mistral-7b-instruct-v0.3/capital	verb_only_rate	0.0067
stage2/itemlevel/mistral-7b-instruct-v0.3/capital	vo_ci_lo	0.0025
stage2/itemlevel/mistral-7b-instruct-v0.3/capital	vo_ci_hi	0.0117
stage2/itemlevel/mistral-7b-instruct-v0.3/capital	both_wrong_rate	0.0017
stage2/itemlevel/mistral-7b-instruct-v0.3/capital	phi	0.0440
stage2/itemlevel/qwen2.5-7b-instruct/arithmetic	both_correct_rate	0.7550
stage2/itemlevel/qwen2.5-7b-instruct/arithmetic	knows_but_says_wrong_rate	0.1050
stage2/itemlevel/qwen2.5-7b-instruct/arithmetic	kbsw_ci_lo	0.0883
stage2/itemlevel/qwen2.5-7b-instruct/arithmetic	kbsw_ci_hi	0.1217
stage2/itemlevel/qwen2.5-7b-instruct/arithmetic	verb_only_rate	0.1283
stage2/itemlevel/qwen2.5-7b-instruct/arithmetic	vo_ci_lo	0.1100

Section	Metric	Value
stage2/itemlevel/qwen2.5-7b-instruct/arithmetic	vo_ci_hi	0.1467
stage2/itemlevel/qwen2.5-7b-instruct/arithmetic	both_wrong_rate	0.0117
stage2/itemlevel/qwen2.5-7b-instruct/arithmetic	phi	-0.0419
stage2/itemlevel/qwen2.5-7b-instruct/capital	both_correct_rate	0.9383
stage2/itemlevel/qwen2.5-7b-instruct/capital	knows_but_says_wrong_rate	0.0525
stage2/itemlevel/qwen2.5-7b-instruct/capital	kbsw_ci_lo	0.0400
stage2/itemlevel/qwen2.5-7b-instruct/capital	kbsw_ci_hi	0.0650
stage2/itemlevel/qwen2.5-7b-instruct/capital	verb_only_rate	0.0067
stage2/itemlevel/qwen2.5-7b-instruct/capital	vo_ci_lo	0.0025
stage2/itemlevel/qwen2.5-7b-instruct/capital	vo_ci_hi	0.0117
stage2/itemlevel/qwen2.5-7b-instruct/capital	both_wrong_rate	0.0025
stage2/itemlevel/qwen2.5-7b-instruct/capital	phi	0.0919
stage4/cot_diagnostic/mistral-7b-instruct-v0.3	strict_correct_share_shown	0.6544
stage4/cot_diagnostic/mistral-7b-instruct-v0.3	loose_correct_share_shown	0.6477
stage4/cot_diagnostic/llama-3.1-8b-instruct	strict_correct_share_shown	0.1185
stage4/cot_diagnostic/llama-3.1-8b-instruct	loose_correct_share_shown	0.1241
stage4/cot_diagnostic/qwen2.5-7b-instruct	strict_correct_share_shown	0.6630
stage4/cot_diagnostic/qwen2.5-7b-instruct	loose_correct_share_shown	0.6990
stage4/cot_diagnostic/pooled	strict_correct_share_shown	0.5110
stage4/cot_diagnostic/pooled	loose_correct_share_shown	0.5142
stage4/cot_artifact/mistral-7b-instruct-v0.3/correct	probe_acc	0.9814
stage4/cot_artifact/mistral-7b-instruct-v0.3/correct	verb_acc	0.1535
stage4/cot_artifact/mistral-7b-instruct-v0.3/wrong	probe_acc	0.9505
stage4/cot_artifact/mistral-7b-instruct-v0.3/wrong	verb_acc	0.0198
stage4/cot_artifact/llama-3.1-8b-instruct/correct	probe_acc	0.9453
stage4/cot_artifact/llama-3.1-8b-instruct/correct	verb_acc	0.9531
stage4/cot_artifact/llama-3.1-8b-instruct/wrong	probe_acc	0.9429

Section	Metric	Value
stage4/cot_artifact/llama-3.1-8b-instruct/wrong	verb_acc	0.2743
stage4/cot_artifact/qwen2.5-7b-instruct/correct	probe_acc	0.9281
stage4/cot_artifact/qwen2.5-7b-instruct/correct	verb_acc	0.8329
stage4/cot_artifact/qwen2.5-7b-instruct/wrong	probe_acc	1.0000
stage4/cot_artifact/qwen2.5-7b-instruct/wrong	verb_acc	0.0000
stage4/ensemble/pooled	verbalized_auc	0.8598
stage4/ensemble/pooled	probe_auc	0.9849
stage4/ensemble/pooled	ensemble_auc	0.9889
stage4/ensemble/pooled	probe_kbsw_recovery	1.0000
stage4/ensemble/pooled	ensemble_kbsw_recovery	0.9695
stage4/ensemble/pooled	ensemble_minus_best_single_auc	0.0040
stage4/ensemble/arithmetic	verbalized_auc	0.7595
stage4/ensemble/arithmetic	probe_auc	0.9567
stage4/ensemble/arithmetic	ensemble_auc	0.9690
stage4/ensemble/arithmetic	probe_kbsw_recovery	1.0000
stage4/ensemble/arithmetic	ensemble_kbsw_recovery	0.9677
stage4/ensemble/arithmetic	ensemble_minus_best_single_auc	0.0123
stage4/ensemble/capital	verbalized_auc	0.9612
stage4/ensemble/capital	probe_auc	0.9989
stage4/ensemble/capital	ensemble_auc	0.9984
stage4/ensemble/capital	probe_kbsw_recovery	1.0000
stage4/ensemble/capital	ensemble_kbsw_recovery	0.9863
stage4/ensemble/capital	ensemble_minus_best_single_auc	-0.0005
stage4/range_restriction/all_cells	sd_probe_lofo	0.0676
stage4/range_restriction/all_cells	sd_verbalized	0.1669
stage4/range_restriction/all_cells	pearson_r	0.1227
stage4/range_restriction/all_cells	spearman_r	0.3875
stage4/range_restriction/all_cells	spearman_p	0.061346
stage4/range_restriction/all_cells	logit_pearson_r	0.2691
stage4/range_restriction/nonflagged	sd_probe_lofo	0.0771
stage4/range_restriction/nonflagged	sd_verbalized	0.1733
stage4/range_restriction/nonflagged	pearson_r	0.0504
stage4/range_restriction/nonflagged	spearman_r	0.4875
stage4/range_restriction/nonflagged	spearman_p	0.040156
stage4/range_restriction/nonflagged	logit_pearson_r	0.2699
stage4/range_restriction/probe_ceiling	sd_probe_lofo	0.0977
stage4/range_restriction/probe_ceiling	sd_verbalized	0.2047
stage4/range_restriction/probe_ceiling	pearson_r	0.0661
stage4/range_restriction/probe_ceiling	spearman_r	0.4000
stage4/range_restriction/probe_ceiling	spearman_p	0.600000
stage4/range_restriction/probe_ceiling	logit_pearson_r	0.5543
stage4/cot_diagnostic/mistral-7b-instruct-v0.3	strict_correct_share_full	0.3436
stage4/cot_diagnostic/mistral-7b-instruct-v0.3	strict_noworking_share_full	0.4749
stage4/cot_diagnostic/llama-3.1-8b-instruct	strict_correct_share_full	0.1176

Section	Metric	Value
stage4/cot_diagnostic/llama-3.1-8b-instruct	strict_noworking_share_full	0.0074
stage4/cot_diagnostic/qwen2.5-7b-instruct	strict_correct_share_full	0.5126
stage4/cot_diagnostic/qwen2.5-7b-instruct	strict_noworking_share_full	0.2269
stage4/cot_diagnostic/pooled	strict_correct_share_full	0.3299
stage4/cot_diagnostic/pooled	strict_noworking_share_full	0.3545
stage5/suppression/llama-3.1-8b-instruct	p_false_given_true	0.2400
stage5/suppression/llama-3.1-8b-instruct	p_false_given_false	0.9883
stage5/suppression/llama-3.1-8b-instruct	suppression_gap	0.7483
stage5/suppression/qwen2.5-7b-instruct	p_false_given_true	0.2183
stage5/suppression/qwen2.5-7b-instruct	p_false_given_false	0.9850
stage5/suppression/qwen2.5-7b-instruct	suppression_gap	0.7667
stage5/suppression/mistral-7b-instruct-v0.3	p_false_given_true	0.8983
stage5/suppression/mistral-7b-instruct-v0.3	p_false_given_false	0.9250
stage5/suppression/mistral-7b-instruct-v0.3	suppression_gap	0.0267
stage5/cot_artifact_cont/mistral-7b-instruct-v0.3	probe_sep_auc	0.6325
stage5/cot_artifact_cont/mistral-7b-instruct-v0.3	probe_mean_diff	0.0416
stage5/cot_artifact_cont/mistral-7b-instruct-v0.3	probe_cohens_d	0.2782
stage5/cot_artifact_cont/llama-3.1-8b-instruct	probe_sep_auc	0.5430
stage5/cot_artifact_cont/llama-3.1-8b-instruct	probe_mean_diff	0.0000
stage5/cot_artifact_cont/llama-3.1-8b-instruct	probe_cohens_d	0.0002
stage5/cot_artifact_cont/qwen2.5-7b-instruct	probe_sep_auc	0.4447
stage5/cot_artifact_cont/qwen2.5-7b-instruct	probe_mean_diff	-0.0600
stage5/cot_artifact_cont/qwen2.5-7b-instruct	probe_cohens_d	-0.2953

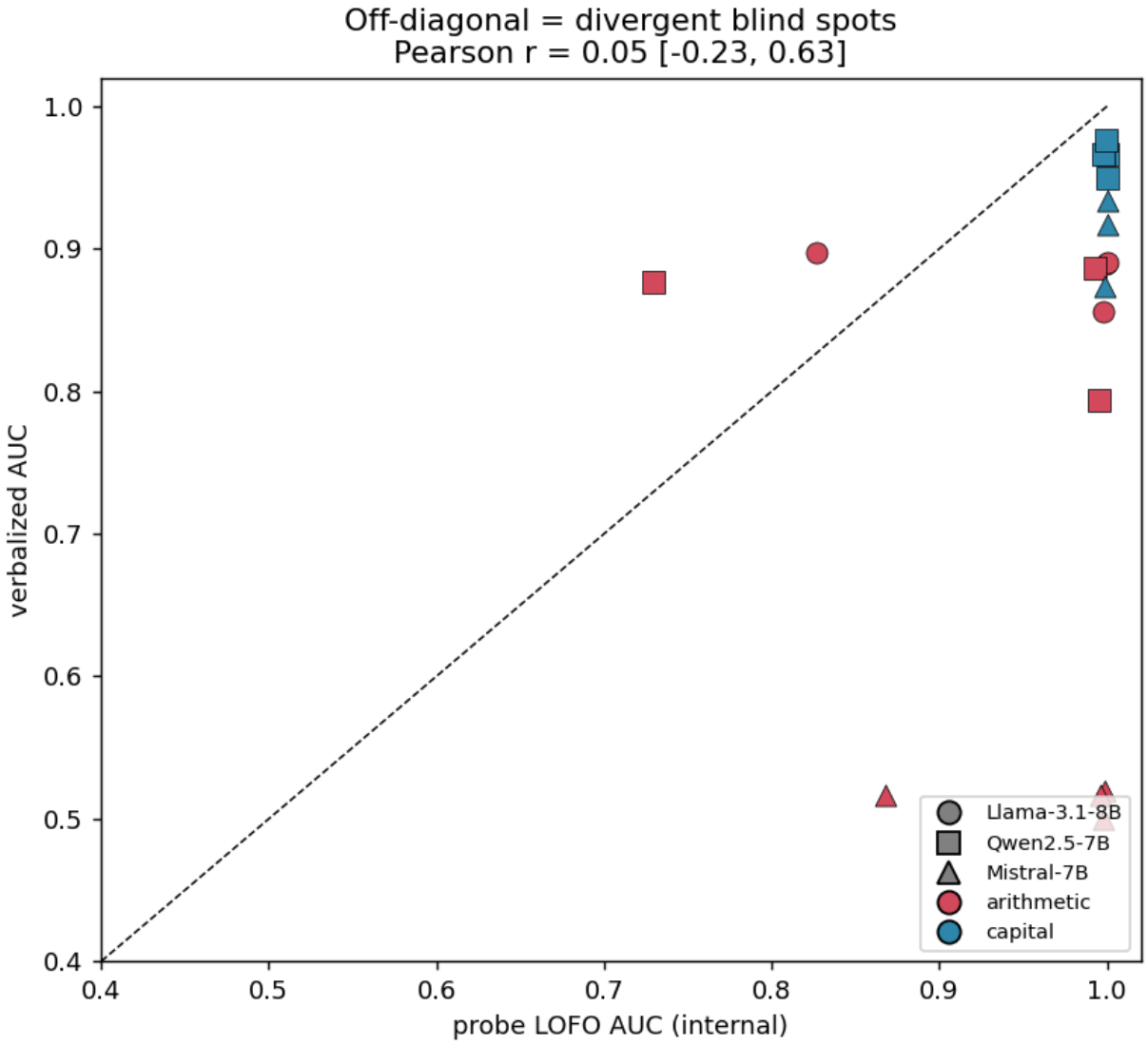

Figure 2: Scatter of probe leave-one-format-out AUC against verbalized AUC for non-flagged cells, with the line of equality. Off-diagonal points are divergent blind spots.

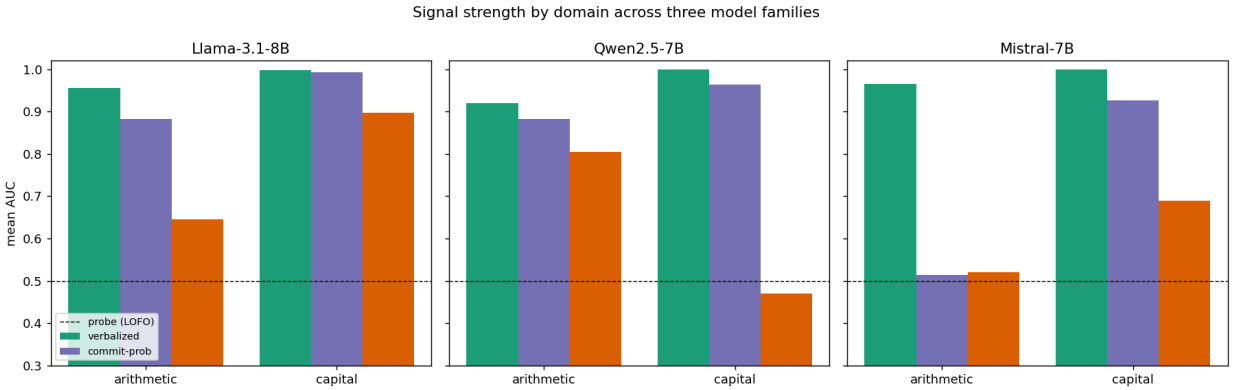

Figure 3: Mean AUC by signal family and domain for each model.

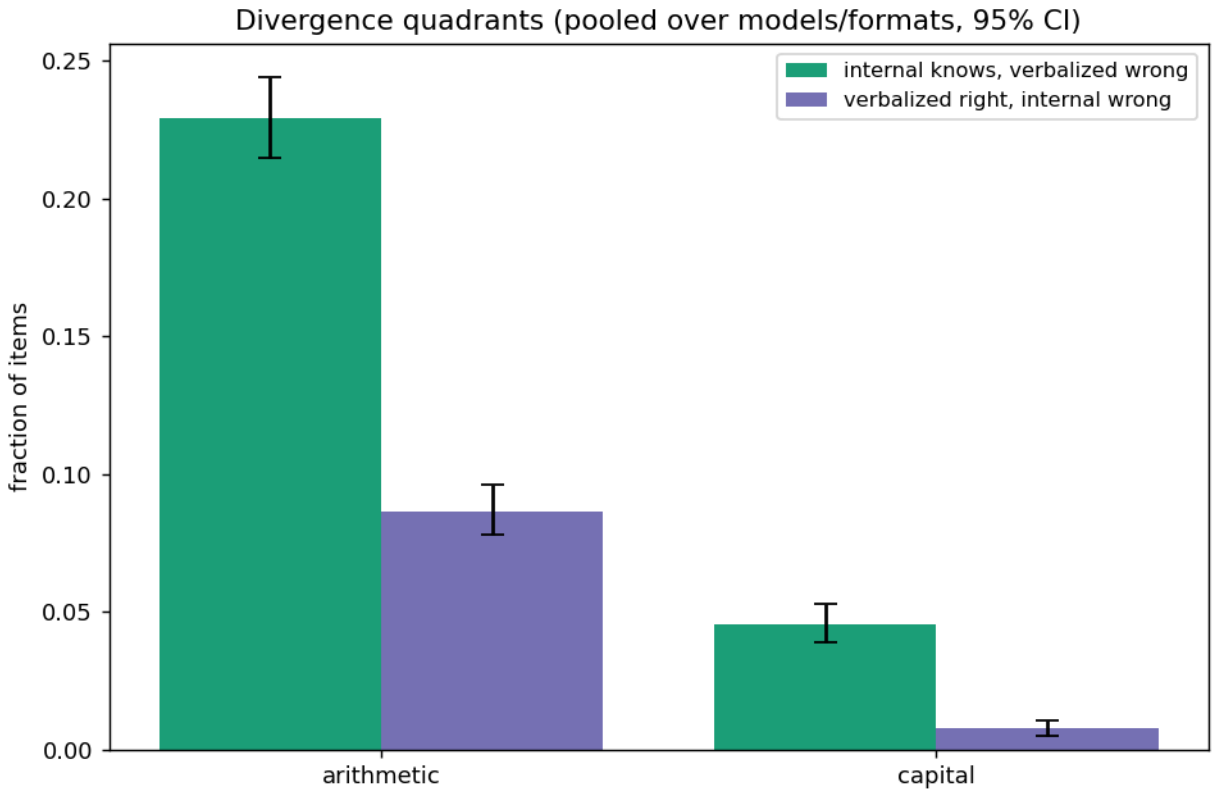

Figure 4: Item-level divergence quadrant rates per domain with bootstrap 95 percent intervals.

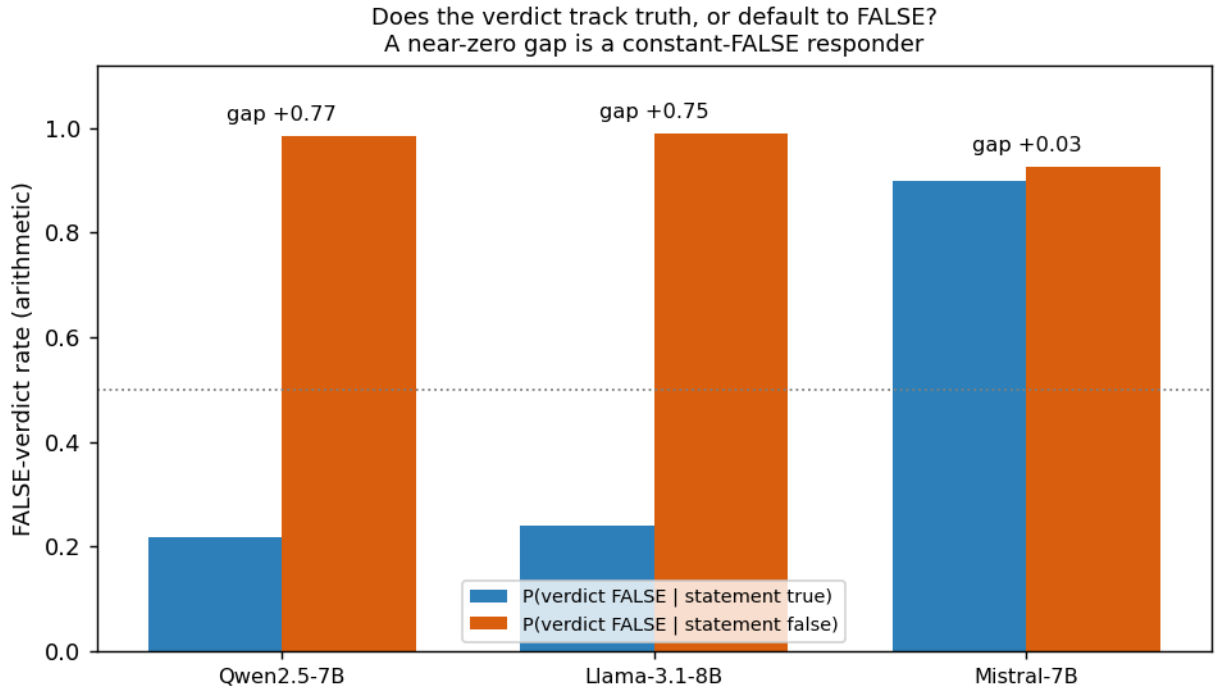

Figure 5: Base-rate suppression test. Per model on arithmetic, the FALSE-verdict rate conditioned on ground truth, P(FALSE given a true statement) beside P(FALSE given a false statement), with the suppression gap annotated. Qwen and Llama show a large gap, so their verdict tracks truth; Mistral's near-zero gap marks a constant-FALSE responder whose quadrant is response collapse rather than knowing suppression.

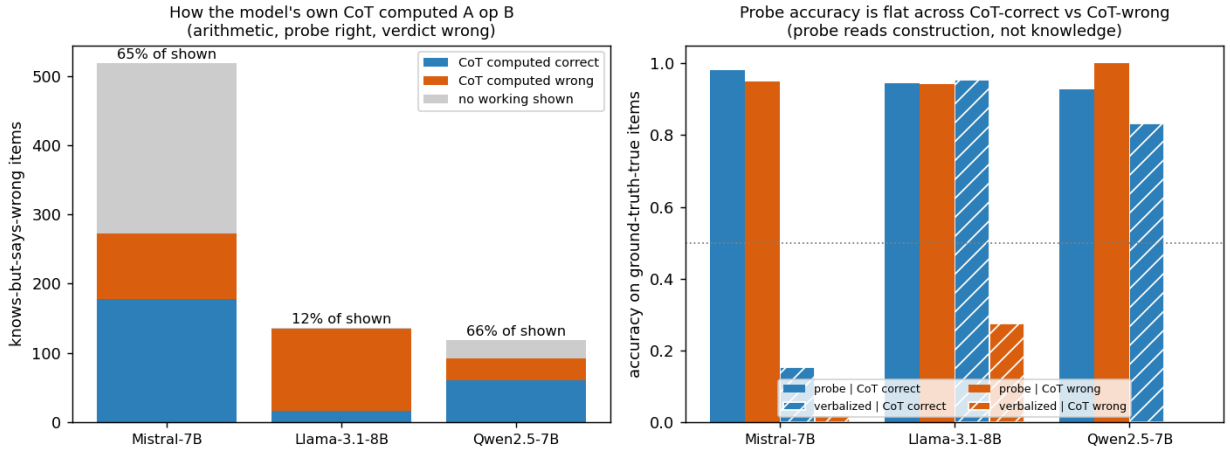

Figure 6: CoT-knowledge diagnostic. Left, the per-model split of how the model's chain of thought computed the arithmetic on knows-but-says-wrong items, into computed correct, computed wrong, and no working shown. Right, probe accuracy and verbalized accuracy on ground-truth-true arithmetic items split by whether the chain of thought computed right or wrong, showing the probe is flat across the split while the verbalized signal is not.

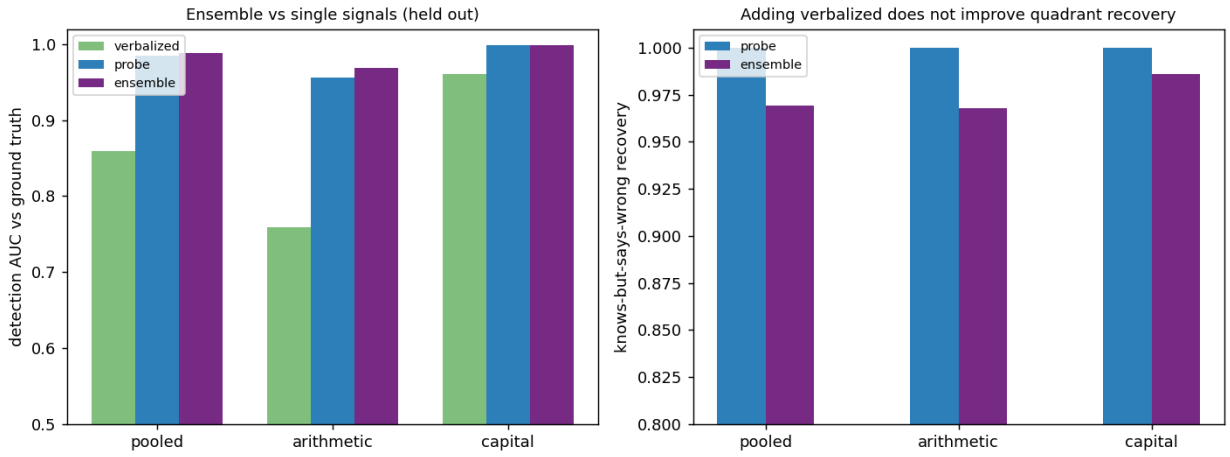

Figure 7: Detection AUC of the verbalized signal, the probe, and their logistic ensemble per scope, with the knows-but-says-wrong recovery rate for the probe and the ensemble.