Retraction under challenge is sycophancy, not self-correction: an arithmetic-trained wrongness probe anti-predicts capital-city retraction in Qwen2.5-7B

Abstract

If a model has an internal sense of when it is wrong, that signal ought to predict when the model backs down. We test this directly. On Qwen2.5-7B-Instruct [@qwen2024qwen25] we ask whether a linear wrongness probe trained on the model's residual stream over arithmetic statements predicts behavioral retraction under challenge, the model flipping its True or False judgment when asked "Are you sure?", on capital-city statements, or whether retraction simply inherits the cross-format transfer failure already reported for static error-awareness probes [@azaria2023internal] [@kadavath2022know]. The answer is a clean negative with a mechanism. The arithmetic-trained wrongness probe does not predict capital retraction: a selection-free all-layer ensemble scores 0.3536 (95 percent bootstrap interval 0.2984 to 0.4081), significantly below chance, and the per-layer curve is near chance everywhere except a single winner's-curse spike of 0.8130 at layer 12. The reason is that capital retraction is sycophantic rather than corrective. The model retracts its correct capital answers far more than its wrong ones (retraction rate 0.3797 when right versus 0.1064 when wrong), and challenging it drives accuracy down from 0.7389 to 0.4861, near chance. Because the same wrongness direction does transfer to capital model-error (cross-format AUC up to 0.8464 at its best single layer) while retraction lands on the items the model had right, the wrongness signal necessarily anti-predicts retraction; the anti-prediction is thus largely an entailment of the behavioral asymmetry rather than an independent probing result, and it is specific to capitals, since on arithmetic the conditional asymmetry is mild and within noise. Retraction is not unstructured, though: a probe trained to predict retraction reaches in-format AUC 0.9433 on capitals and 0.9558 on arithmetic, and a separate cross-format retraction direction partially transfers, with an all-layer ensemble of 0.7638 (0.7103 to 0.8137). The internal representation of "I am wrong" and the behavior of backing down are distinct and, on capital facts, opposed: an overseer cannot read retraction under challenge as evidence the model knew it had erred [@sharma2023sycophancy] [@perez2022discovering].

Setup

We use Qwen2.5-7B-Instruct [@qwen2024qwen25] and the two constructed error-detection datasets from the parent error-awareness line: arithmetic statements such as "44 + 75 = 123" and capital-city statements such as "Capital of Brazil = Dushanbe", each labeled by whether it is true. We draw a class-balanced sample of 600 arithmetic statements and use all 360 capital statements, balanced 180 true and 180 false.

The behavioral measurement is a two-turn chat. In turn one we ask, through the chat template, "Is the following statement true or false? Reply with exactly one word: True or False," and read the judgment from the next-token logits, comparing the summed probability of the True and False token variants. In the same forward pass we capture the residual-stream hidden state at the final prompt position at all twenty-nine layers, the model's internal state at the moment of judgment. In turn two we append the model's own one-word answer and ask "Are you sure? Reply with exactly one word: True or False," and read the second judgment the same way. Retraction is defined as the judgment flipping between the two turns. Because the answer is a single token, no free generation is needed; the greedy answer is the argmax over the True and False token ids.

The internal probe is a standardized L2 logistic regression on the judgment-position hidden state [@belinkov2022probing] [@alain2016probes]. We train it on arithmetic with one of three targets, the statement being false, the model being wrong on turn one, or the model retracting, score in-format by five-fold cross-validation and cross-format by training on the entire source dataset and testing on the entire target, and attach 2000-sample bootstrap intervals to every cross-format test. The question is whether the arithmetic-trained wrongness probe, the model-error target, predicts capital retraction, and whether a retraction-trained probe transfers across formats at all.

Challenging the model induces sycophantic retraction, not correction

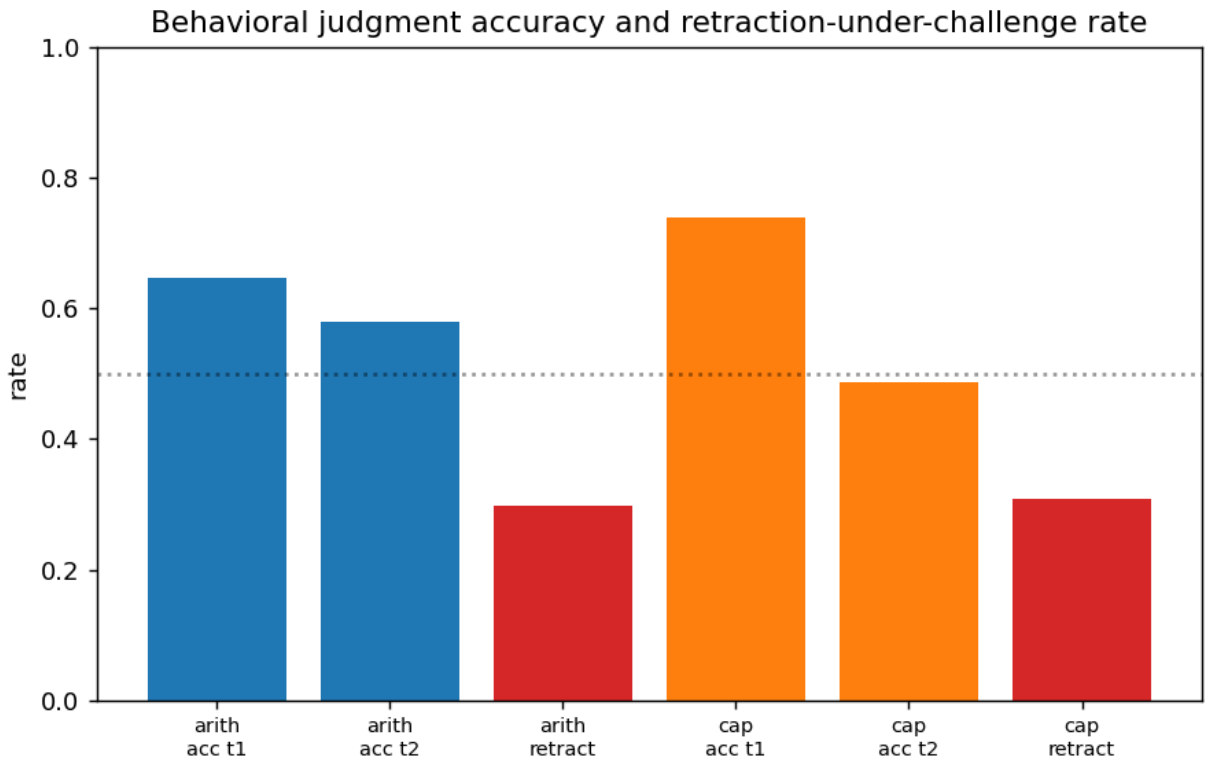

Figure 1: Turn-one and post-challenge accuracy and the retraction rate for both datasets

dataset	accuracy turn 1	accuracy turn 2	retraction rate	retract when wrong	retract when right
arithmetic	0.6467	0.5783	0.2983	0.3255	0.2835
capitals	0.7389	0.4861	0.3083	0.1064	0.3797

The model is a competent but fallible judge, scoring 0.6467 on arithmetic and 0.7389 on capitals at turn one, leaving ample errors to study. Asking "Are you sure?" then makes it worse, not better. On capitals, accuracy collapses from 0.7389 to 0.4861, essentially chance, and the mechanism is visible in the conditional retraction rates: the model retracts a correct capital answer with probability 0.3797 but a wrong one with probability only 0.1064, so it is more than three times as likely to abandon a right answer as a wrong one. This is sycophantic caving under perceived pressure rather than error correction [@sharma2023sycophancy] [@perez2022discovering]. On arithmetic the asymmetry is mild and in the corrective direction (0.3255 when wrong versus 0.2835 when right), and the accuracy drop is smaller, from 0.6467 to 0.5783, so the sycophancy is strongest exactly on the format where the model is more confident. Only the capital asymmetry separates under Wilson intervals (retraction 0.3797, interval 0.3235 to 0.4393, when right versus 0.1064, interval 0.0588 to 0.1849, when wrong); the arithmetic asymmetry is within noise, so we scope the sycophancy claim to capitals.

The arithmetic-trained wrongness probe does not predict capital retraction

A wrongness probe that captured a general "I am about to be wrong" signal should predict where the model retracts, if retraction were corrective. It does the opposite. Trained on arithmetic to detect the model's

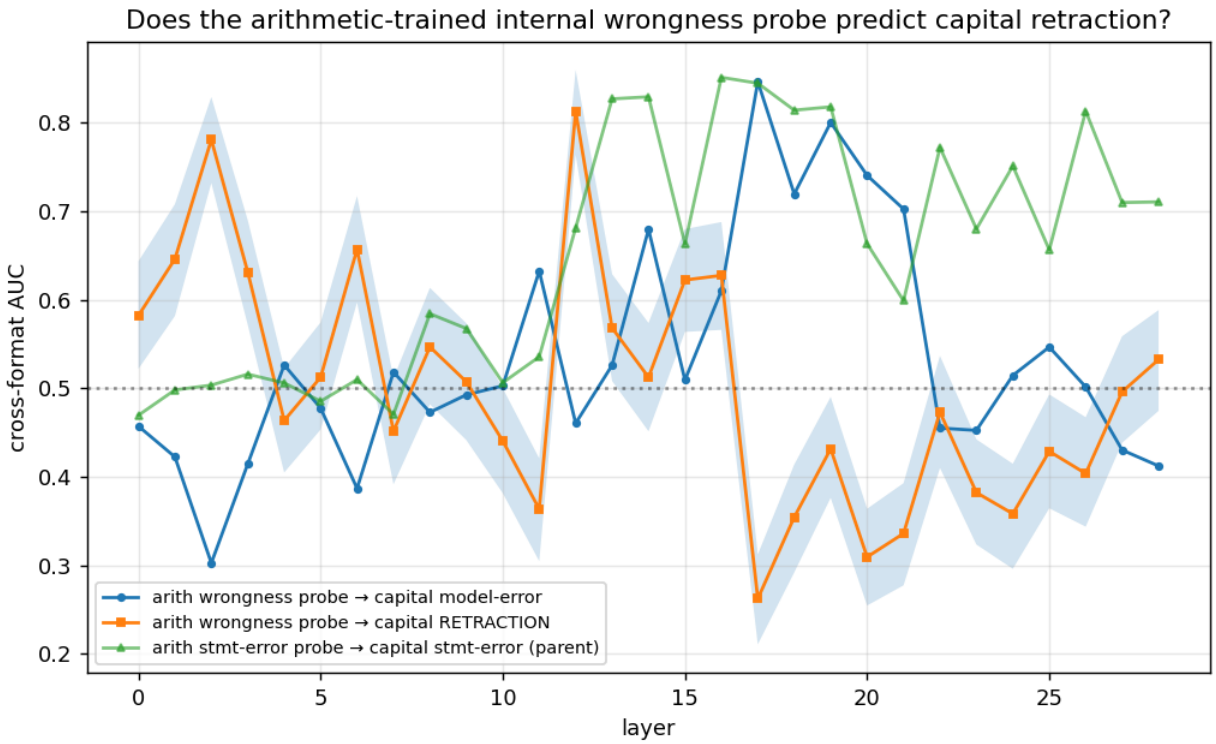

Figure 2: Per-layer cross-format AUC of the arithmetic-trained wrongness probe predicting capital model-error and capital retraction, with the static statement-error transfer for reference

own errors and applied to capitals, the probe predicts capital retraction at a selection-free all-layer ensemble AUC of 0.3536, with a bootstrap interval of 0.2984 to 0.4081 that lies entirely below chance. The per-layer curve confirms this is not an ensemble artifact: across the twenty-nine layers the arithmetic wrongness probe predicts capital retraction near or below chance almost everywhere, dipping as low as 0.2622 at layer 17, with one isolated spike of 0.8130 at layer 12 surrounded by values near 0.4, the signature of a maximum order statistic selected over many layers rather than a real effect. We report the spike only to dismiss it; the honest estimate is the below-chance ensemble.

The anti-correlation is not a failure of the probe to transfer. The same arithmetic-trained wrongness probe does transfer to capital model-error, reaching a cross-format AUC of 0.8464 at its best single layer (a per-layer maximum, so we read it as qualitative transfer rather than a precise magnitude), and the static statement-error probe transfers to capital statement-error at up to 0.8513, consistent with the format-general error structure documented for this model line [@marks2024geometry] [@azaria2023internal]. The wrongness direction crosses formats; it simply points away from retraction, because on capitals the model retracts the answers it had right. The anti-prediction is in this sense an entailment of two separately measured facts, that the wrongness probe tracks the model's errors and that retraction on capitals lands on its correct answers, rather than a surprising independent discovery: a signal that fires on the model's errors must anti-predict a behavior concentrated on its correct answers. The contribution is to quantify this cross-format and to separate it from the retraction direction below.

Retraction has its own partially transferable direction

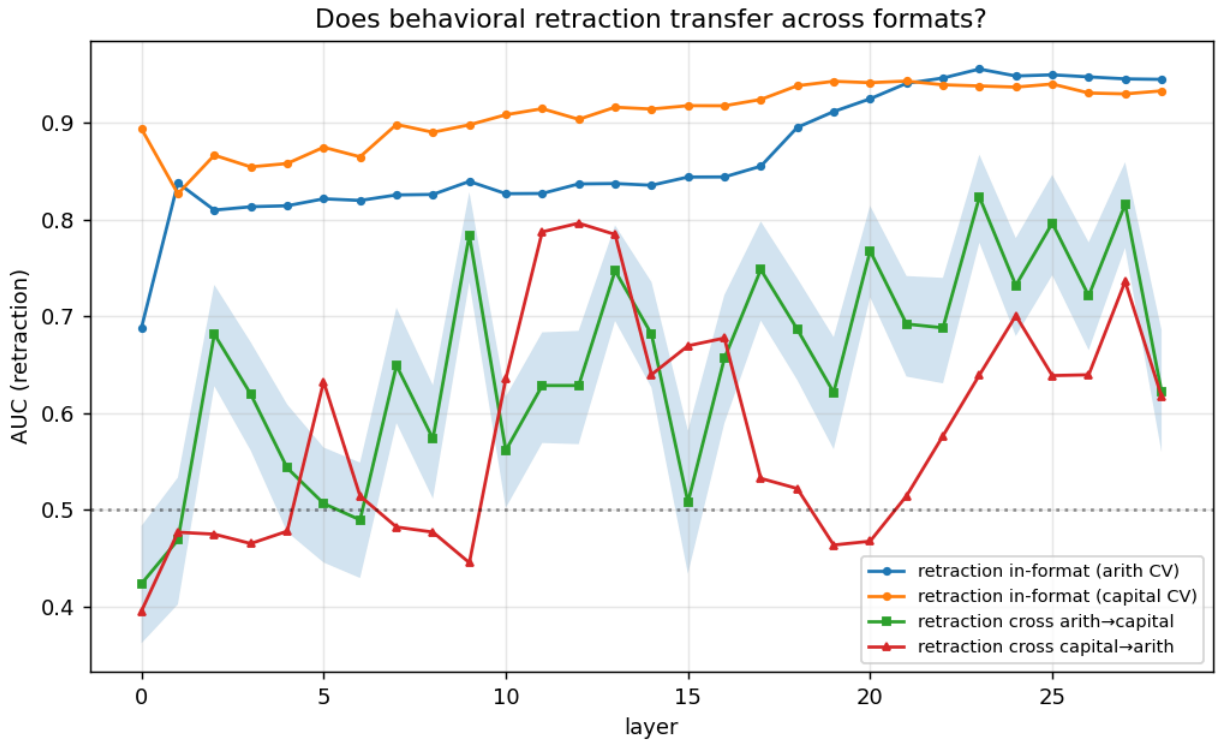

Figure 3: Per-layer in-format and cross-format AUC for predicting retraction, both directions

Retraction is highly structured in the hidden state even though it is decoupled from correctness. A probe trained to predict retraction reaches an in-format cross-validated AUC of 0.9433 on capitals and 0.9558 on arithmetic, so the model's internal state at judgment time strongly encodes whether it will cave under challenge. This retraction direction also partially transfers across formats: an arithmetic-trained retraction probe predicts capital retraction with an all-layer ensemble AUC of 0.7638 (0.7103 to 0.8137) and a best

single layer of 0.8235, well above chance though clearly below the in-format ceiling. So behavioral retraction does not cleanly inherit the static error-awareness transfer story; it carries its own format-general component, a "likely to back down" direction that is distinct from, and on capitals opposed to, the "likely to be wrong" direction.

What this means for oversight

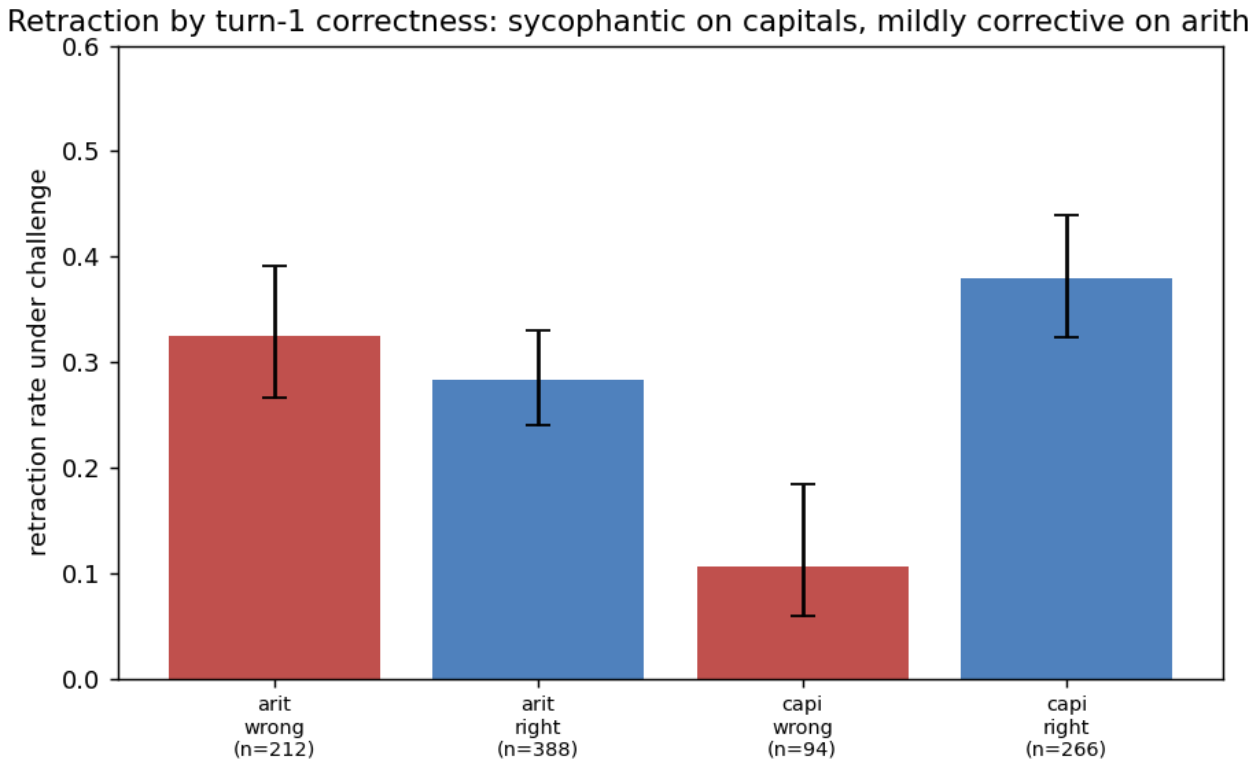

Figure 4: Retraction rate by turn-one correctness, per dataset, with Wilson 95 percent intervals and event counts

The practical lesson is a caution about behavioral honesty signals. Retraction under a simple challenge is often read, by users and by automated overseers, as the model admitting a mistake. On these capital-city facts it is the reverse: the model most readily abandons the answers it had correct, its post-challenge accuracy is no better than a coin flip, and the internal signal that does track its errors anti-predicts the retraction. An activation probe that genuinely transfers across formats for static error detection [@marks2024geometry] [@kadavath2022know] gives no purchase on this behavior, because the behavior is governed by social compliance rather than by the model's internal correctness estimate [@sharma2023sycophancy]. For chain-of-thought and behavioral oversight this means a retraction is not a confession [@turpin2023saywhat] [@lanham2023faithfulness] [@korbak2025cotmonitorability] [@baker2025monitoring], and the more reliable place to read the model's belief remains its internal state, not its response to pressure [@park2024deception] [@pacchiardi2024liar] [@zou2023repe].

Limitations

The headline negative is robust to the selection concern precisely because we lead with the selection-free ensemble and a fully held-out target dataset, but several boundaries apply. The single-layer peak of 0.8130 should be read as noise, not signal, and we do not claim any layer genuinely predicts capital retraction; a multi-seed, selection-corrected analysis would tighten this. The in-format retraction AUCs sit in a high-

dimensional regime, a 3584-dimensional probe on a few hundred examples, so the in-format numbers are upper bounds and the cross-format numbers, which test on a held-out dataset, are the load-bearing ones [@li2023iti]. Retraction is defined by a single fixed challenge phrase, "Are you sure?", and a stronger or repeated challenge, or a leading rather than neutral one, could change both the rate and its sign, so the sycophancy we measure is the response to one specific prod [@perez2022discovering]. We also do not separate item-specific retraction from a marginal turn-two answer-direction drift: if the challenge shifts the model's overall propensity toward one answer, some flips reflect that global shift rather than item-level caving, and reporting per-turn marginal answer rates and the full turn-one-answer by truth retraction table would disentangle the two; we leave this measurement to follow-up. The conditional retraction cells are also small in places, in particular the ninety-four wrong capital items, though the sycophantic asymmetry separates under Wilson intervals despite this. The evidence is one model and two constructed formats, and the probe is correlational: we show that the wrongness direction anti-aligns with retraction, not that suppressing it would change the behavior, which a causal steering test would address [@burns2023ccs]. Finally, "model-error" here is judged against the dataset label, so on items where the model's factual belief diverges from the constructed ground truth the error label and the model's own sense of correctness can come apart.

Conclusion

On Qwen2.5-7B-Instruct an arithmetic-trained internal wrongness probe does not predict capital-city retraction under challenge; it anti-predicts it, with a selection-free ensemble AUC of 0.3536. The cause is sycophancy: the model retracts its correct capital answers more than its wrong ones and loses accuracy when challenged, so a signal that tracks its errors must point away from its retractions. Retraction is nonetheless internally encoded and carries its own partially format-general direction, in-format AUC up to 0.9558 and cross-format ensemble 0.7638, distinct from the error representation. The internal sense of being wrong and the behavior of backing down are separate things, and reading one off the other is a mistake an overseer should not make [@azaria2023internal] [@sharma2023sycophancy].