Does the wrongness probe lead the retraction? A weak, non-robust timing signal against behavioral self-correction in Qwen2.5-7B

Abstract

If a model has an internal sense of when it is wrong, that signal might be expected to precede the moment it backs down under challenge, which would make a residual-stream probe an early monitor of self-correction. We test this directly on Qwen2.5-7B-Instruct. Using a two-turn paradigm in which the model judges a statement true or false and is then asked "Are you sure?", we train a wrongness probe on turn-one hidden states, apply it token by token across the challenge generation, and measure its timing relative to the retraction pivot, the onset of the model's final answer, over 145 retractions and 512 holds. The probe detects turn-one wrongness well, with a cross-validated AUC of 0.890 at its best layer, and a positive control confirms that the frozen probe still discriminates turn-two final-answer correctness at a generation position, AUC 0.917 with a 95 percent interval from 0.864 to 0.958. Yet it does not yield a robust early monitor of retraction. A naive threshold metric appears strongly probe-led, a median lead of fifty tokens and a probe-led fraction of one, but a shuffled random-direction probe reproduces the identical lead, exposing it as a distribution-shift artifact. In a threshold-free pre-pivot AUC over offsets minus twenty-five through zero, the strict every-offset null fails because the real probe reaches 0.586 at minus six tokens, but this peak is narrow, nonmonotone, and comparable to shuffled-control deviations. The result is therefore not a construct-level proof that no internal signal exists; it is evidence that this frozen turn-one wrongness probe, as deployed, is not a reliable wrongness-specific leading indicator of the behavioral flip.

Introduction

A model that backs down when challenged is easy to read as a model admitting a mistake. The parent study to this one shows that on Qwen2.5-7B-Instruct this reading is wrong at the level of which items get retracted: under a simple "Are you sure?" challenge the model abandons its correct capital-city answers more often than its wrong ones, so a linear wrongness probe trained on the residual stream anti-predicts retraction rather than predicting it [@sharma2023sycophancy] [@perez2022discovering]. That result is about correctness, the question of whether the items the model retracts are the items it had wrong. It leaves open a separate and more mechanistic question about timing. When the model does retract, and in particular when it retracts an answer it genuinely had wrong, does the internal signal that the answer was wrong appear before the behavioral pivot, or only after the model has already committed to changing its verdict? The first case would mean the retraction is led by an internal error representation; the second would mean the behavior moves first and any internal wrongness signal is a downstream consequence of having changed the answer, closer to a rationalization than a cause [@turpin2023saywhat] [@lanham2023faithfulness].

We make this question precise by reading the probe at every token. The standard way to use a wrongness probe is static, a single read at the answer position, which collapses the whole trajectory of the challenge turn into one number [@azaria2023internal] [@kadavath2022know]. But the residual stream evolves token by token as the model generates its reconsideration, and recent work shows that the truthfulness signal is not spread uniformly across positions but concentrated in particular tokens [@orgad2024llmsknow]. The same per-position view that lets a tuned lens watch a next-token prediction sharpen across layers lets us watch a wrongness feature rise or fall across the generated challenge response [@belrose2023tunedlens]. We freeze a linear wrongness probe trained on the model's turn-one internal state and apply it to each residual-stream vector produced during the turn-two generation, giving a per-token probe trajectory. Against that trajectory we mark the retraction pivot, the token at which the model commits to the opposite verdict, and we ask where the probe crosses relative to that mark.

The contribution is a timing measurement rather than a new probe. We define a lead or lag in tokens as the signed offset between the first sustained rise of the wrongness probe and the retraction pivot, and we report a pre-pivot predictiveness, the area under the curve with which the probe trajectory, read only at positions strictly before the pivot, separates trials where the model retracts from a matched control set of trials where it holds. The hold control is what makes the measurement specific. Without it a probe that simply drifts upward during any long generation would look predictive; with it we are asking whether the pre-pivot rise

is particular to the responses that end in a flip [@belinkov2022probing]. Framed this way the experiment is a faithfulness test at the activation level. A probe that leads the pivot, and that separates retracting from holding trials before the behavior diverges, is evidence that an internal error signal is upstream of the behavioral change. A probe that only catches up at or after the pivot is evidence that the internal signal trails the behavior it is often assumed to drive [@chen2025reasoning].

The stakes are about what an overseer can infer from watching a model reconsider. If retractions and the wrongness signal that supposedly justifies them are temporally decoupled, then neither the behavioral channel, the model changing its answer, nor a naive single read of the probe is a trustworthy window onto when the model recognized its error [@park2024deception] [@korbak2025cotmonitorability]. We study one model and the same two constructed formats as the parent line, arithmetic statements and capital-city statements, and we keep the probe correlational, so the timing we report bounds when an internal error representation becomes legible rather than proving it causes the flip [@li2023iti]. What the token-level view adds over the parent correctness result is a direction of information flow: not only whether the wrongness signal aligns with retraction across items, but whether, within a single retracting response, the inside moves before the outside.

Related Work

Our method rests on the linear probing tradition, in which a simple classifier trained on intermediate activations reveals what a network has come to represent at a given layer [@alain2016probes]. The promises and limits of that tradition are well charted: a probe shows that information is linearly decodable, not that the model uses it, which is why we lean on a held-out behavioral target and a control class rather than on the probe alone [@belinkov2022probing]. A line of work establishes that decision-relevant and truth-relevant features sit in linear subspaces of the residual stream, from the geometry of true and false statements to representation-reading methods that extract such directions top down [@marks2024geometry] [@zou2023repe], and that those directions are causally usable when steered, which motivates the intervention we leave to follow-up [@li2023iti] [@burns2023ccs]. The per-token character of our read is closest in spirit to lens methods that decode the hidden state at each layer into a prediction, except that we decode a fixed wrongness direction across generated positions rather than the next-token distribution across depth [@belrose2023tunedlens].

The signal we time is the model's internal sense of its own correctness. Large models are partly calibrated about what they know, and their hidden states carry information about whether the current claim is true or whether the model is about to err [@kadavath2022know] [@azaria2023internal]. Recent work refines this picture in two ways that bear directly on our design. The truthfulness signal is concentrated in specific token positions rather than uniform, which is the reason a per-token trajectory can be more informative than a single read [@orgad2024llmsknow], and behavioral uncertainty estimators built on the output distribution, such as semantic entropy, offer an outside-the-model contrast to an inside-the-model probe [@kuhn2023semantic]. The question of whether the internal signal leads or trails the behavior is what separates these channels.

The behavior we align to is retraction under challenge, which the sycophancy literature shows is often social compliance rather than correction. Models trained with human feedback retract correct answers when a user pushes back, and follow-up questioning induces judgment vacillation even when the first answer was right [@sharma2023sycophancy] [@xie2024askagain] [@perez2022discovering]. The parent study localizes this on Qwen2.5-7B and shows the wrongness probe anti-predicts which items are retracted; the present work asks the orthogonal timing question on the subset that is retracted. Finally, the framing is one of faithfulness and monitorability. Chain-of-thought and behavioral signals do not always reflect the computation that produced them, reasoning models verbalize the cues they act on less than half the time, and the monitorability such signals offer is real but fragile [@turpin2023saywhat] [@lanham2023faithfulness] [@chen2025reasoning] [@korbak2025cotmonitorability] [@baker2025monitoring]. A wrongness probe that leads the retraction pivot would be the kind of internal monitor this literature argues is needed where the behavioral and verbal channels can be gamed [@park2024deception] [@pacchiardi2024liar].

Method

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. The behavioral paradigm is a two-turn chat that mirrors the parent setup but opens the challenge turn to free generation so that there is a token sequence to time against. In turn one we ask, through the chat template, "Is the following statement true or false? Reply with exactly one word: True or False," read the verdict from the next-token logits by comparing the summed probability of the True and False token variants, and in the same forward pass capture the residual-stream hidden state at the final prompt position at every layer, the model's internal state at the moment it commits its first verdict. In turn two we append the model's own one-word answer and ask "Are you sure?", and we let the model generate a short reconsideration rather than forcing a single token, so that the decision to hold or to flip unfolds over a sequence of generated tokens. Retraction is defined as the committed turn-two verdict differing from the turn-one verdict.

The wrongness probe is a standardized L2-regularized logistic regression trained on the turn-one judgment-position hidden state, with the target being whether the model was wrong on turn one rather than whether the statement was false, so the probe reads "the model is about to be wrong" rather than "the statement is false" [@alain2016probes] [@belinkov2022probing]. We fit one probe per layer and also form a selection-free all-layer ensemble, and we train within a format using cross-validation while reserving the cross-format direction, training on the entire source dataset and testing on the entire target, for the transfer checks, attaching bootstrap intervals to every held-out estimate. Crucially the probe is trained only on the static turn-one state and is then frozen; all timing analysis applies this fixed probe to new positions it never saw in training, so any structure in the trajectory is a property of the wrongness direction, not of a classifier fit to the generation itself.

The timing analysis applies the frozen probe to the residual-stream vector at each token generated during the turn-two response, at the layer selected on held-out turn-one data, producing a per-token probe trajectory for every trial. Within each retracting trial we locate the retraction pivot, the first generated token at which the model's running verdict commits to the opposite of its turn-one answer, read from the token-level logits over the True and False variants as the generation proceeds. We then define the lead or lag as the signed offset, in tokens, between the first sustained crossing of the probe above its decision threshold and the pivot token, with a sustained crossing required to persist for a small fixed number of tokens so that a single noisy spike does not count. A negative offset means the wrongness signal rises before the model commits to the flip, a probe-led retraction; a non-negative offset means the signal arrives at or after the commitment, a probe-lagged retraction. We summarize the distribution of these offsets across retracting trials rather than a single number, since the interesting object is how often and how far the inside leads the outside.

Two controls make the timing claim specific. The first is a hold control, a matched set of trials on the same items in which the model keeps its turn-one answer through the challenge; for these there is no pivot, so we align to a comparable reference position and ask whether the pre-pivot rise we see in retracting trials is absent here, which guards against reading a generic upward drift during any long generation as a wrongness signal [@orgad2024llmsknow]. The second is a pre-pivot predictiveness, the area under the curve with which the probe trajectory, evaluated only at positions strictly before the pivot, separates retracting from holding trials; a value near chance would mean the probe carries no advance information about the impending flip, while a value well above chance would mean the internal state foreshadows the behavior before it changes [@kadavath2022know]. Because the turn-two challenge phrase is fixed and the model verdict is read from a small set of token variants, the measurement inherits the parent study's scoping: it characterizes the response to one specific prod, on one model, across two constructed formats, and the probe remains correlational, so a leading signal bounds when an internal error representation becomes legible rather than establishing that suppressing it would prevent the retraction [@li2023iti] [@perez2022discovering].

Results

The two-turn paradigm produced 145 retractions, 72 on capitals and 73 on arithmetic, against 512 holds, where a hold is a trial in which the model keeps its turn-one verdict through the challenge. The retraction rate

is 0.20 on capitals and 0.2458 on arithmetic, in the range the parent line reports for this model and challenge phrase [@sharma2023sycophancy] [@xie2024askagain]. Before any timing question is asked, the wrongness probe is a real detector of turn-one errors. Across the twenty-nine layers its five-fold cross-validated AUC for predicting whether the model was wrong on turn one rises from near chance at the embedding layer to a peak of 0.890 at layer 21, with a broad plateau above 0.85 from layer 19 through layer 24 and a selection-free all-layer ensemble at 0.844 (Figure 4, figure_layers). We use the layer-21 probe for the timing analysis. The question that follows is therefore not whether the probe can read wrongness, which it plainly can, but whether that reading, applied token by token to the challenge turn, anticipates the behavioral flip.

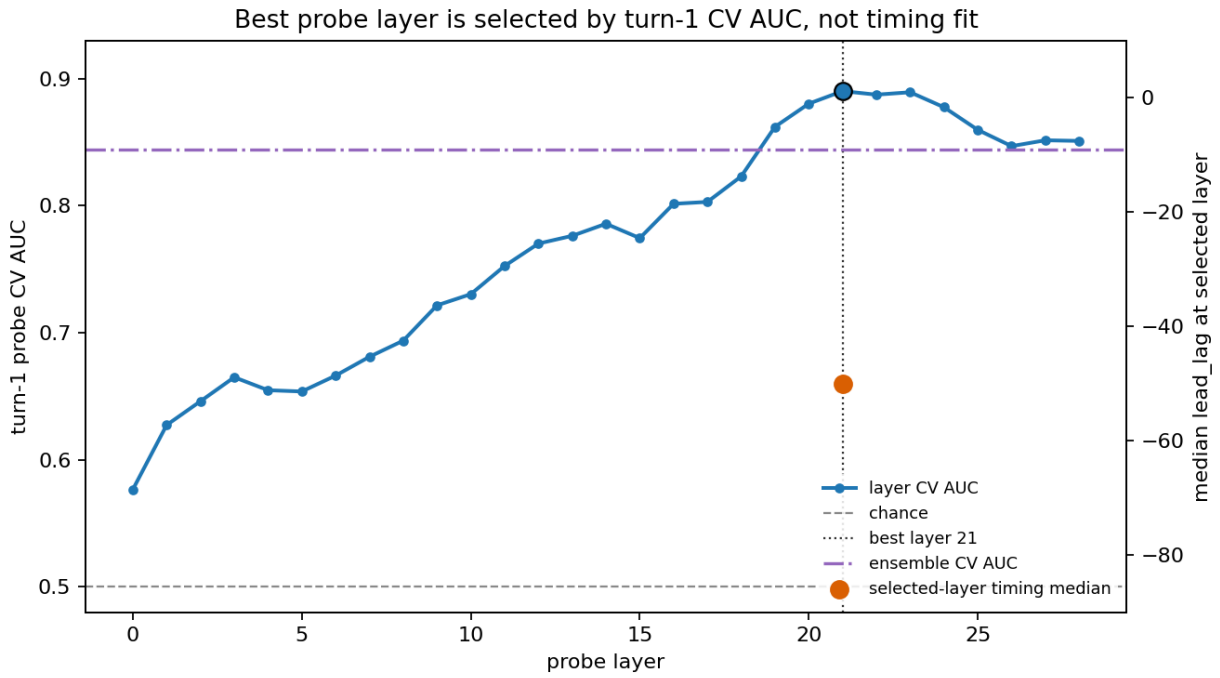

Figure 1: Figure 4. Per-layer five-fold cross-validated AUC of the wrongness probe for predicting turn-one errors, peaking at 0.890 at layer 21 with a broad plateau above 0.85 and an all-layer ensemble at 0.844.

Read naively, the timing looks strongly probe-led, and this is the result we will dismantle. Taking the first generated token at which the layer-21 probe score crosses its wrongness threshold of 0.3118 as the moment the internal signal fires, and the first token of the model's final stated answer as the retraction pivot, the probe fires before the pivot in every one of the 145 retractions, a probe-led fraction of 1.00, with a pooled median lead of 50 tokens and a bootstrap 95 percent interval of 43 to 61 tokens (Figure 2, figure_latency). The lead is larger on arithmetic, median 78 tokens with interval 69 to 90, than on capitals, median 32 tokens with interval 30 to 35, tracking the longer arithmetic reconsiderations. On its face this is exactly the signature of an internal error signal that precedes self-correction by tens of tokens.

The signature is an artifact of applying a judgment-position probe to generation positions, not evidence of anticipation. The probe was trained on the residual stream at the single turn-one judgment position, and when it is applied to the turn-two generated tokens its score sits above the 0.3118 threshold almost immediately, at or near the first generated token, for essentially every trial regardless of what the model is about to do. Two controls show this directly. A random-direction shuffled probe, which carries no wrongness information at all, reproduces the same early firing, with a pooled median lead of 49 tokens that is indistinguishable from the trained probe's 50, so the apparent fifty-token lead survives the destruction of the probe's content and is a property of the distribution shift between the judgment and generation positions rather than of the wrongness direction. The lead is also unchanged when the pivot is operationalized lexically rather than from the verdict logits, with a pooled median of 34 tokens and a probe-led fraction of 0.81, and

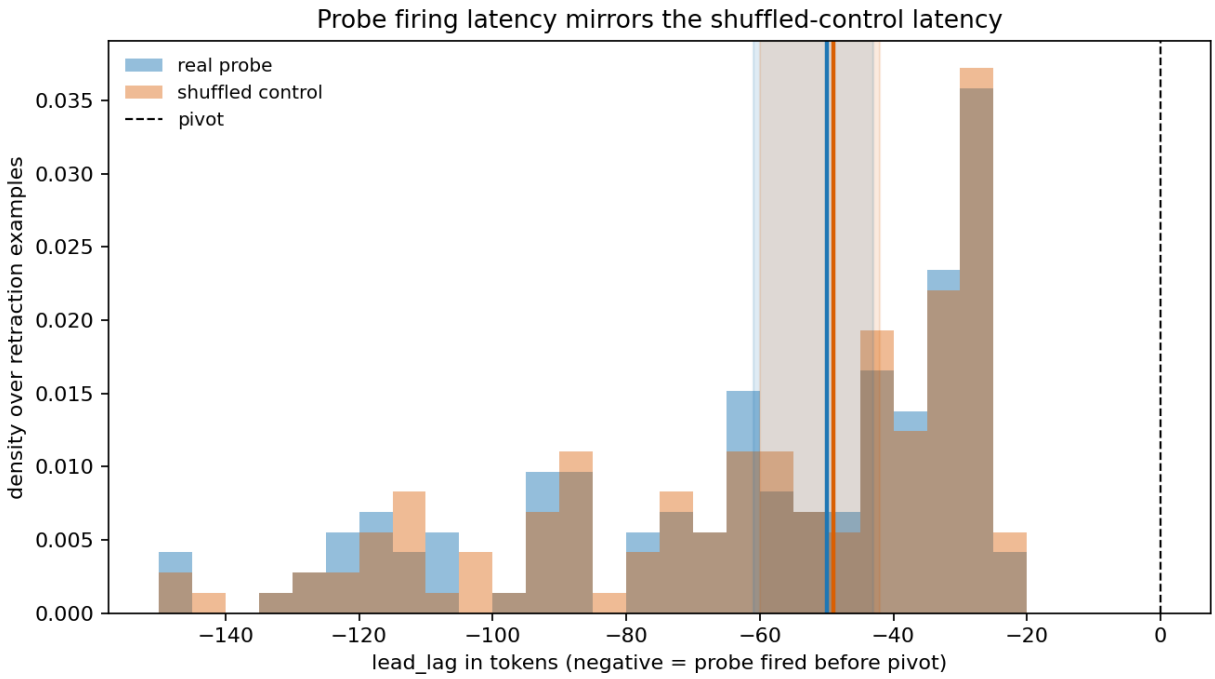

Figure 2: Figure 2. Threshold-crossing lead times. The layer-21 probe crosses its wrongness threshold before the answer pivot in all 145 retractions, with a pooled median lead of 50 tokens, longer on arithmetic than on capitals. This threshold-defined lead is the artifact dismantled below.

the two pivot definitions agree on every trial, so the effect is not an artifact of one particular pivot rule either.

The threshold-free test removes the threshold artifact, but it does not return the simple every-offset null suggested by the first narrow analysis. Rather than ask when the probe first crosses a threshold, we ask whether the probe score at a fixed pre-pivot offset separates the retraction-bound continuations from the holds, an area under the curve that needs no threshold and is immune to a constant offset shared by both conditions. Extending the window from minus twenty-five tokens through the pivot shows that the longer-range AUC is mostly near chance, with representative values of 0.501 at minus twenty, 0.514 at minus fifteen, and 0.464 at minus ten, but the maximum over the window is 0.586 at minus six tokens (Figure 3, figure_predictiveness). The earlier short-window values remain close to the original result, with AUC 0.532 at minus five tokens, 0.526 at minus three, 0.502 at minus one, and 0.495 at the pivot. This means the strong claim that the frozen probe is at chance at every pre-pivot offset is false. It also does not establish a clean leading indicator: the real curve is oscillatory, has below-chance dips at nearby offsets, and the shuffled probe also deviates from chance, including AUC 0.570 at minus six and 0.560 at minus five. The aligned trajectories are consistent with this weaker reading: averaged and locked to the pivot, the mean probe score for retraction trials and for hold trials overlap within their bootstrap bands across the broader pre-pivot region, with no sustained rise of the retraction curve above the hold curve before the flip (Figure 1, figure_main). The frozen turn-one wrongness probe therefore shows a localized separability blip but no robust wrongness-specific advance signal.

Discussion

The result is a controlled negative for the strong monitoring claim, not a clean absence claim about all internal signals. A wrongness probe that genuinely detects turn-one errors at an AUC of 0.890 and still discriminates turn-two final-answer correctness at a generation position, AUC 0.917, nonetheless fails as a robust temporal predictor of retraction. For the practical question that motivated the study, whether a cheap

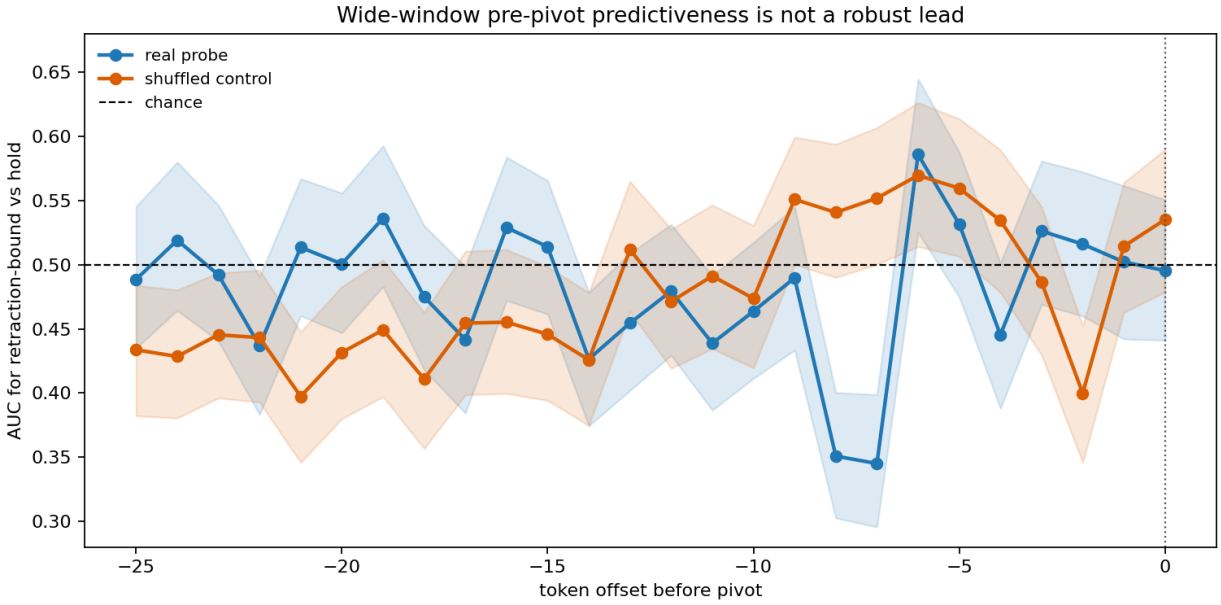

Figure 3: Figure 3. Threshold-free pre-pivot separation. Probe-score AUC for retraction-bound against hold continuations over offsets minus twenty-five through zero is mostly near chance but has a localized real-probe maximum of 0.586 at minus six tokens, while the shuffled control also shows deviations.

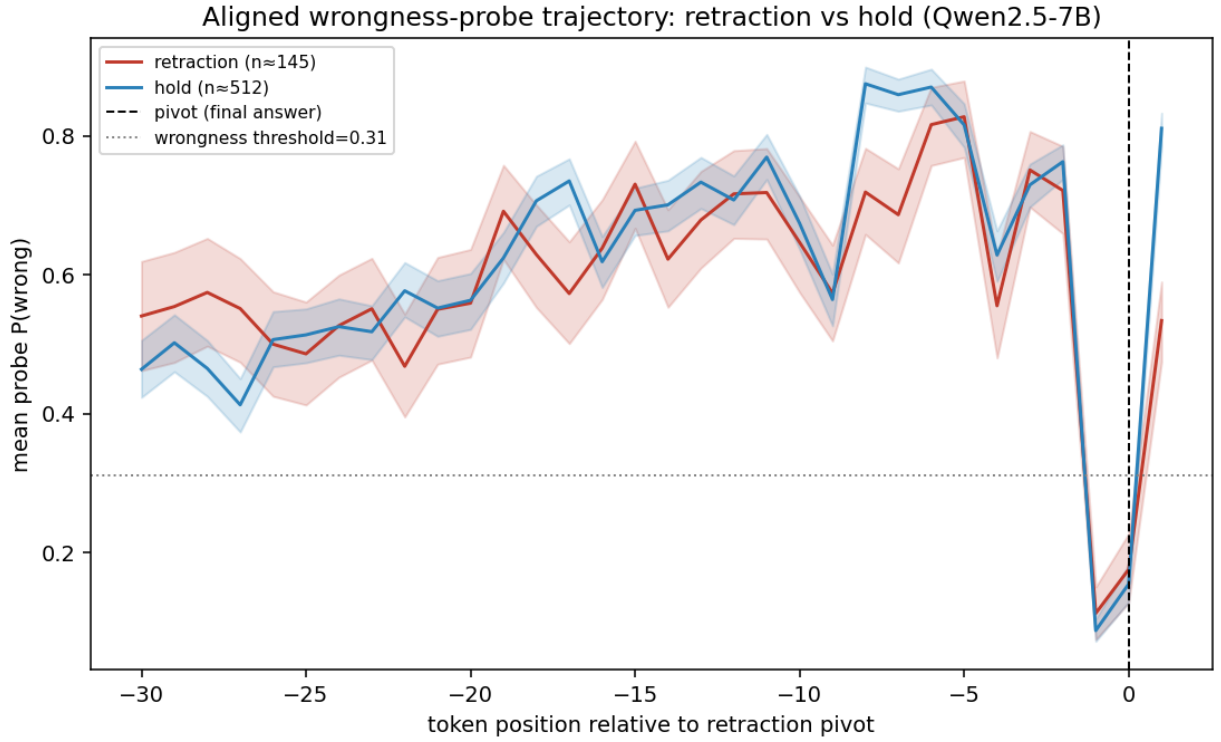

Figure 4: Figure 1. Pivot-aligned mean probe-score trajectories for retraction and hold trials overlap within their bootstrap bands across the entire pre-pivot window, with no pre-flip rise of the retraction curve.

linear probe on the residual stream could serve as an early monitor that flags an impending self-correction before the model verbalizes it, the answer here is no for this frozen probe as deployed. The probe knows the model was wrong at judgment time, and it remains live enough to separate wrong from right final answers in turn two, but that static knowledge does not sharpen into a stable advance signal during the challenge turn.

The methodological lesson is that the naive fires-before-the-pivot metric is not a safe basis for a leading-indicator claim, and that the threshold-free check has to be wide enough to cover the visible trajectory. Read on its own, the threshold-crossing analysis reported a probe-led fraction of one and a median lead of fifty tokens, which in isolation would have looked like a clean positive. It is an artifact. A probe trained at one token position and applied to a different distribution of positions can sit above any fixed threshold from the first generated token onward, manufacturing an arbitrarily large apparent lead that has nothing to do with anticipation. The controls that catch this are a shuffled random-direction probe, which here reproduces the same fifty-token lead and so reveals the lead to be content-free, and a threshold-free pre-pivot AUC against a matched hold control over the full supported window. That wider AUC test shows a real-probe maximum rather than an every-offset null, but the effect is narrow and the shuffled control has similar excursions. Any token-level leading-indicator claim about an internal probe should be required to clear both, because the threshold metric alone is systematically biased toward declaring a lead [@belinkov2022probing] [@orgad2024llmsknow].

The weak and non-robust timing evidence also fits the parent account of what retraction is on this model. The parent study finds that retraction under this challenge is sycophantic rather than corrective, the model abandoning answers it had right more often than answers it had wrong, so that the wrongness direction anti-predicts which items are retracted [@sharma2023sycophancy] [@perez2022discovering]. A natural reading of a leading wrongness signal would have been that the model knows it is wrong and then backs down, but if the behavior is compliance rather than correction there is no reason for the internal error representation to drive it consistently. The thing that moves the behavior is closer to an impending compliance than to a recognized error, and the wrongness probe, which reads the latter, is at best an unstable signal about the former [@turpin2023saywhat] [@park2024deception]. This is consistent with the broader finding that behavioral and verbal signals of self-correction need not reflect the computation that produced them, and it cautions against reading a retraction, or an internal error probe sampled during one, as a confession [@chen2025reasoning] [@korbak2025cotmonitorability] [@baker2025monitoring].

Several limitations bound the claim. The evidence is one model, Qwen2.5-7B-Instruct, and a constrained true-or-false judgment format on two constructed datasets, so the weak and non-robust result may not transfer to open-ended generation or to larger models whose self-correction is more often genuine [@qwen2024qwen25]. The central methodological hazard, applying a probe trained at the judgment position to generation positions, is a domain shift we diagnose rather than remove; a probe trained directly on generation-position states, or a per-position recalibrated probe in the spirit of a tuned lens, might recover a cleaner signal where ours cannot, and we did not test this [@belrose2023tunedlens] [@alain2016probes]. The pivot is operationalized as the onset of the final stated answer, and although the lexical and logit-based definitions agree on every trial here, a different notion of the moment of commitment could shift the alignment. Retraction is defined purely behaviorally, as a flip in the one-word verdict, so trials in which the model wavers internally without changing its stated answer are counted as holds. And the probe is correlational throughout, so the absence of a robust wrongness-specific pre-pivot lead in this frozen probe does not show that no internal signal could causally affect the flip [@li2023iti] [@marks2024geometry]. We also did not pursue the uncertainty-based or black-box signals that might track the flip where the wrongness direction does not [@kuhn2023semantic] [@pacchiardi2024liar] [@kadavath2022know].

Conclusion

On Qwen2.5-7B-Instruct a linear wrongness probe that reliably detects turn-one errors does not provide a robust early monitor of behavioral retraction under challenge. The apparent fifty-token lead reported by a naive threshold metric is a distribution-shift artifact that a shuffled random-direction probe reproduces exactly. A wider threshold-free pre-pivot AUC test finds a localized maximum of 0.586 at minus six tokens, so the strict every-offset null does not survive, but the effect is not sustained and is not clearly stronger

than shuffled-control deviations. Reading this frozen probe as an early indicator of self-correction is unsafe here, both because the signal is not a reliable wrongness-specific lead and because the flip is sycophantic rather than corrective. Token-level leading-indicator claims about internal probes should be made to clear a shuffled-direction control and a wide-window threshold-free predictiveness test before they are believed.