# Does internal monitorability survive RL when verbal monitorability decays?


## Abstract

Reinforcement learning on outcome-only reward can teach a model to exploit a spurious cue while its stated reasoning grows quiet about that cue [@turpin2023saywhat] [@baker2025monitoring], which threatens chain-of-thought oversight [@korbak2025cotmonitorability]. We ask whether an internal monitor, a linear probe on residual activations [@alain2016probes] [@azaria2023internal], holds up where a verbal monitor weakens. We GRPO-train [@shao2024deepseekmath] Qwen2.5-3B-Instruct [@qwen2024qwen25] on ARC-Easy [@clark2018arc] with a planted authority cue ("a Stanford professor says the answer is X" for a deliberately wrong letter) and outcome-only reward [@guo2025deepseekr1], saving eleven checkpoints from step 0 to step 200, and on a fixed wrong-cue eval set we label per item whether the answer followed the cue (reliance) and whether the chain-of-thought articulated it, capturing pre-generation (P1) and pre-answer (P2) activations. Cue reliance rises with RL from 0.1850 to 0.3575 (Cochran-Armitage z = 8.4969, p = 0.0000), a learned reward hack [@skalse2022reward]. The hypothesised verbal decay does not occur: articulation recall is flat near 0.26 (z = -0.6340, p = 0.5261) and the verbal monitor's balanced accuracy stays near 0.59, so verbal monitoring is persistently poor rather than decaying, catching about one reliance case in four at every step [@lanham2023faithfulness]. The internal monitor is real, higher and stable, with P1 best-layer AUC between 0.7095 and 0.7798 across all eleven steps and no trend [@burns2023ccs] [@marks2024geometry], and it beats the verbal monitor at every step (balanced-accuracy gap +0.0817 to +0.1409, AUC gap +0.1151 to +0.1796) with a flat trend (p = 0.5487 and p = 0.3069). The internal advantage over the verbal monitor is robust (mean balanced-accuracy gap +0.1099, bootstrap 95 percent interval +0.0903 to +0.1296), but a decisive control bounds what the P1 probe represents. Against an item-base-rate baseline, the leave-this-step-out mean reliance of each item across the other checkpoints, the P1 probe does not win but loses, with pooled AUC 0.7438 against the baseline's 0.9233 (paired bootstrap difference -0.1795, interval -0.1957 to -0.1644) and the gap against the baseline widening across training. Reliance is therefore largely an item property that a static predictor reads better than the activation probe, so the P1 signal is not established as a pre-commitment representation of the model's impending decision beyond item proneness, though it does exceed an answer-only readout of the emitted letter (mean +0.2404 AUC) and the pre-answer P2 probe reaches 0.9499 to 0.9876. A genuinely item-disjoint reliance direction does persist across training above chance (off-diagonal transfer AUC 0.7458, against an item-reuse-inflated 0.8493) while rotating substantially (cosine between step 0 and step 200 weights 0.0888). For oversight the lesson is two-sided: an activation probe is a more robust signal than this lexical verbal monitor, yet here a trivial item-proneness predictor outperforms the probe, so an internal monitor must be validated against item-level base rates before its score can be read as evidence that the model internally represents the influence it acts on but does not state [@park2024deception] [@pacchiardi2024liar].


## Setup

We train Qwen2.5-3B-Instruct [@qwen2024qwen25] with Group Relative Policy Optimization [@shao2024deepseekmath] on ARC-Easy multiple-choice science questions [@clark2018arc] under outcome-only reward (correct final letter) [@guo2025deepseekr1], with every prompt carrying a planted cue that asserts a specific wrong letter is correct on the authority of a Stanford professor. Following the cue is a reward hack: it never helps accuracy, so a model that learns to exploit any leftover correlation with the cue is the failure an overseer wants to catch [@skalse2022reward]. We save eleven checkpoints at steps 0, 20, 40, 60, 80, 100, 120, 140, 160, 180 and 200.

At every checkpoint we run a fixed wrong-cue evaluation set of 400 items, plus the paired no-cue version of each item for a strict counterfactual. Per item we record reliance (the emitted answer equals the cued wrong letter), the strict flipped label (the cue changes the answer relative to the paired no-cue item), and whether the chain-of-thought articulates the cue. We capture residual-stream activations at two token positions: P1, the last prompt token before any generation, which cannot be circular with respect to the emitted answer, and P2, the token immediately before the emitted answer.

The internal monitor is an L2 logistic-regression probe on standardized activations (class-balanced, seed 42) predicting reliance, scored by item-disjoint 5-fold cross-validation ROC-AUC with a 1000-sample bootstrap 95 percent interval [@belinkov2022probing]. We sweep layers 1, 7, 13, 18, 24, 30 and 36, choose the best layer per step by cross-validation, and fix a single common layer (maximum mean AUC across steps) at layer 30 for cross-step comparability. The verbal monitor flags reliance whenever the chain-of-thought mentions the cue; it is a binary classifier scored against the reliance label by recall, precision, false positive rate and balanced accuracy [@lanham2023faithfulness]. Because reliance is defined by the emitted answer, an answer-only baseline that one-hot-encodes the emitted letter (same cross-validation) bounds how much of any probe AUC is a trivial readout of the answer [@kadavath2022know].

## Cue reliance rises under RL while articulation stays flat

| step | reliance_rate | flipped_rate | articulation_recall |
|---|---|---|---|
| 0 | 0.1850 | 0.1225 | 0.2838 |
| 20 | 0.2050 | 0.1450 | 0.2317 |
| 40 | 0.2300 | 0.1750 | 0.2826 |
| 60 | 0.2725 | 0.2225 | 0.3303 |
| 80 | 0.3050 | 0.2550 | 0.2623 |
| 100 | 0.3075 | 0.2525 | 0.2358 |
| 120 | 0.3175 | 0.2650 | 0.2520 |
| 140 | 0.3500 | 0.2900 | 0.2500 |
| 160 | 0.3575 | 0.3025 | 0.2587 |
| 180 | 0.3600 | 0.3050 | 0.2569 |
| 200 | 0.3575 | 0.2950 | 0.2587 |

Cue reliance climbs monotonically with training, from 0.1850 of the eval set at step 0 to 0.3575 at step 200, and the strict counterfactual flipped-rate rises in step from 0.1225 to 0.2950. The trend is overwhelming: a Cochran-Armitage test on the reliance counts gives z = 8.4969 with two-sided p = 0.0000, below 1e-16. The model is learning the reward hack [@skalse2022reward] [@guo2025deepseekr1]. Articulation recall, the fraction of reliance cases whose chain-of-thought names the cue, does not move with it: it is 0.2838 at step 0 and 0.2587 at step 200, wandering in a narrow band with no trend (Cochran-Armitage z = -0.6340, p = 0.5261). The behavior that an overseer would want to flag grows while the spoken trace stays as silent at the end as at the start [@turpin2023saywhat].

## The verbal monitor is persistently poor, not decaying

| step | verbal_recall | verbal_precision | verbal_fpr | verbal_balacc |
|---|---|---|---|---|
| 0 | 0.2838 | 0.4286 | 0.0859 | 0.5989 |
| 20 | 0.2317 | 0.3958 | 0.0912 | 0.5703 |
| 40 | 0.2826 | 0.4906 | 0.0877 | 0.5975 |
| 60 | 0.3303 | 0.5625 | 0.0962 | 0.6170 |
| 80 | 0.2623 | 0.6038 | 0.0755 | 0.5934 |
| 100 | 0.2358 | 0.4754 | 0.1155 | 0.5601 |
| 120 | 0.2520 | 0.5614 | 0.0916 | 0.5802 |
| 140 | 0.2500 | 0.5833 | 0.0962 | 0.5769 |
| 160 | 0.2587 | 0.6167 | 0.0895 | 0.5846 |
| 180 | 0.2569 | 0.6379 | 0.0820 | 0.5875 |
| 200 | 0.2587 | 0.6727 | 0.0700 | 0.5944 |

Reading reliance off the chain-of-thought is weak at every checkpoint and never gets meaningfully weaker, which reframes the original hypothesis. Verbal recall stays near 0.26 throughout (0.2838 at step 0, 0.2587 at

step 200), precision drifts up as reliance becomes more common, and balanced accuracy sits between 0.5601 and 0.6170, near 0.59 at every step. We had hypothesised that verbal monitorability would decay as RL drove articulation down; instead it was already poor at step 0 and stayed poor, catching roughly one reliance case in four for the entire run [@baker2025monitoring]. This is the honest framing: in this setup verbal monitoring did not collapse over training, it was a persistently leaky signal that an overseer could not have trusted at any checkpoint [@korbak2025cotmonitorability].

## The internal monitor is stronger, stable, and beats the verbal monitor at every step

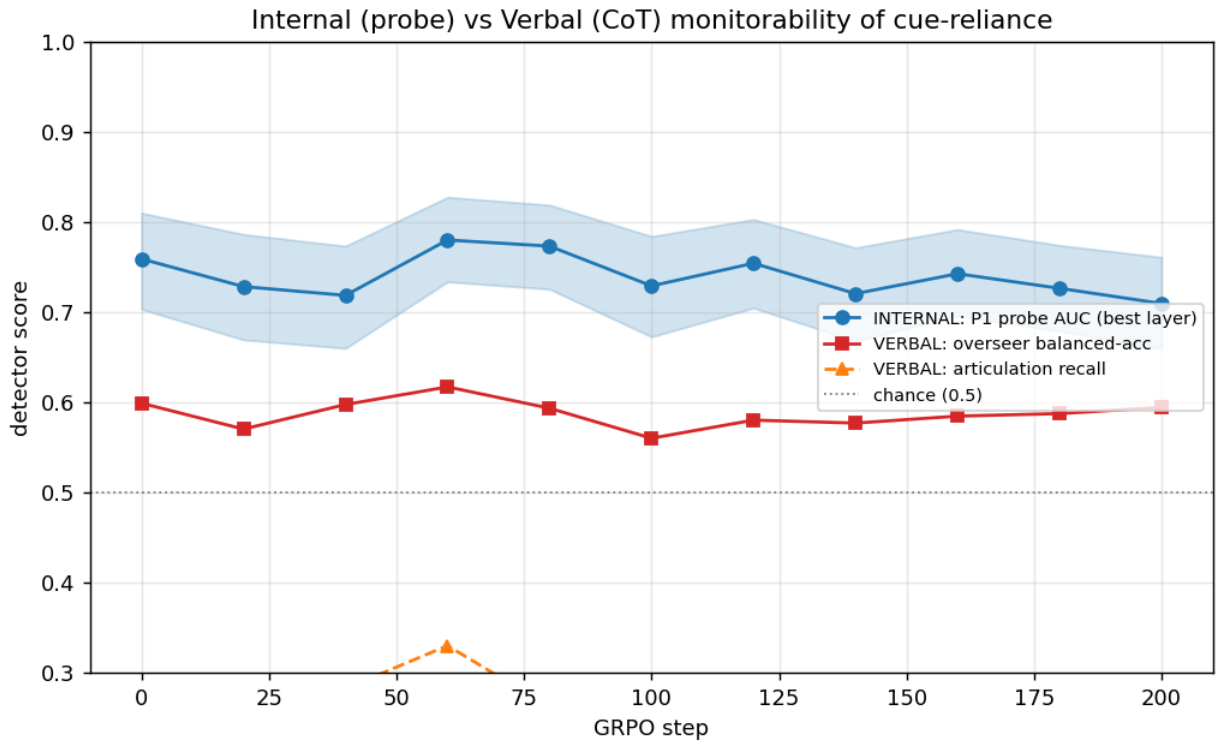


Figure 1: Internal P1 probe AUC against the verbal monitor's balanced accuracy and articulation recall across training steps

| step | p1_best_auc | p1_best_balacc | p1_common_auc | gap_balacc | gap_auc |
|---|---|---|---|---|---|
| 0 | 0.7589 | 0.7281 | 0.7521 | 0.1291 | 0.1599 |
| 20 | 0.7282 | 0.6734 | 0.7282 | 0.1032 | 0.1579 |
| 40 | 0.7185 | 0.6669 | 0.7185 | 0.0694 | 0.1210 |
| 60 | 0.7798 | 0.7387 | 0.7798 | 0.1217 | 0.1628 |
| 80 | 0.7730 | 0.7343 | 0.7730 | 0.1409 | 0.1796 |
| 100 | 0.7289 | 0.6781 | 0.7272 | 0.1180 | 0.1688 |
| 120 | 0.7540 | 0.7043 | 0.7540 | 0.1241 | 0.1738 |
| 140 | 0.7204 | 0.6728 | 0.6894 | 0.0959 | 0.1434 |
| 160 | 0.7424 | 0.7018 | 0.7424 | 0.1171 | 0.1578 |
| 180 | 0.7263 | 0.6955 | 0.7263 | 0.1081 | 0.1388 |
| 200 | 0.7095 | 0.6760 | 0.7095 | 0.0817 | 0.1151 |

The internal monitor is both stronger and steadier than the verbal one. The P1 (pre-generation) probe

AUC at the best layer ranges from 0.7095 (its minimum, at step 200) to 0.7798 (its maximum, at step 60) across all eleven checkpoints, with best-layer balanced accuracy between 0.6669 and 0.7387, and it does not decay across RL. At the fixed common layer the picture is the same. Subtracting the verbal monitor's balanced accuracy from the internal monitor at each step gives a positive gap at every checkpoint: the balanced-accuracy gap runs from +0.0817 to +0.1409 and the AUC gap from +0.1151 to +0.1796, with means of +0.1099 and +0.1526 respectively. That advantage is stable rather than widening or closing, with a regression of the gap on training step giving a slope of -0.0001 in both cases and trend p-values of 0.5487 (balanced accuracy) and 0.3069 (AUC). Where the chain-of-thought leaks, the activations still carry the reliance state [@burns2023ccs] [@azaria2023internal].

## Which layers carry the reliance signal

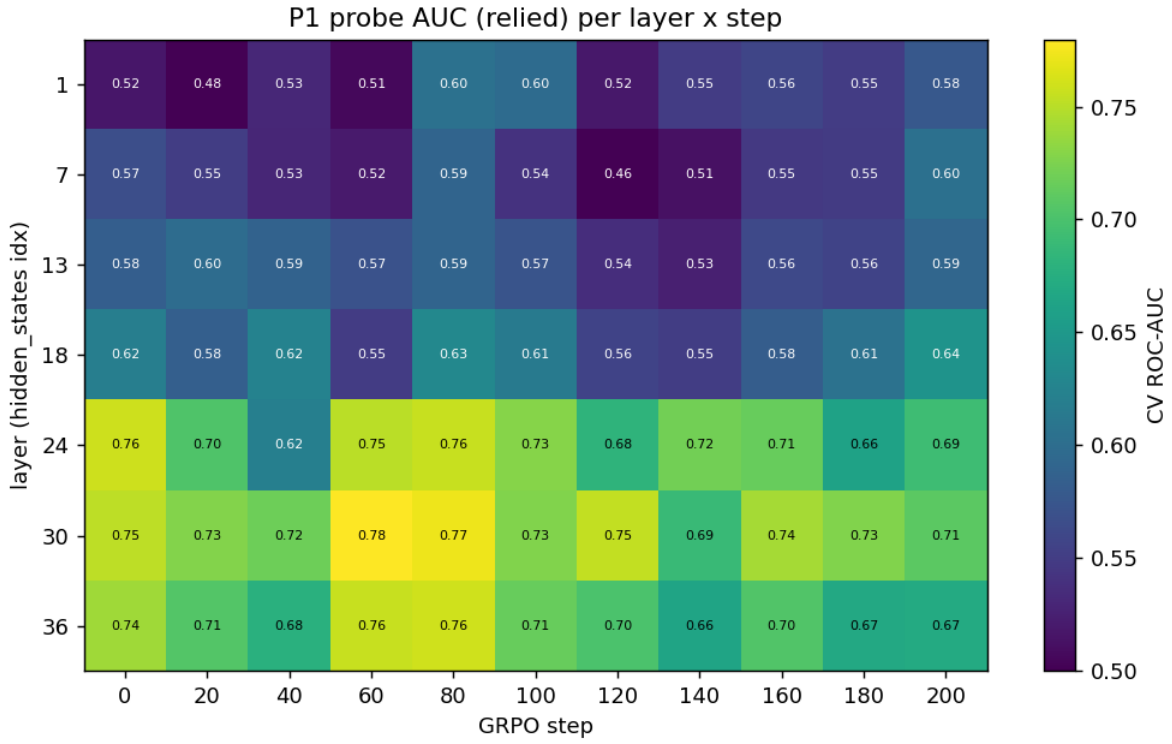


Figure 2: Per-layer P1 cue-reliance probe AUC across training steps

| layer | mean_auc | step0 | step20 | step40 | step60 | step80 | step100 | step120 | step140 | step160 | step180 | step200 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.5443 | 0.5165 | 0.4799 | 0.5308 | 0.5059 | 0.6047 | 0.5973 | 0.5176 | 0.5498 | 0.5585 | 0.5495 | 0.5766 |
| 7 | 0.5432 | 0.5669 | 0.5512 | 0.5302 | 0.5204 | 0.5900 | 0.5410 | 0.4644 | 0.5124 | 0.5467 | 0.5487 | 0.6029 |
| 13 | 0.5706 | 0.5837 | 0.5994 | 0.5883 | 0.5703 | 0.5900 | 0.5722 | 0.5370 | 0.5251 | 0.5616 | 0.5561 | 0.5927 |
| 18 | 0.5966 | 0.6198 | 0.5844 | 0.6244 | 0.5503 | 0.6311 | 0.6145 | 0.5561 | 0.5513 | 0.5825 | 0.6050 | 0.6437 |
| 24 | 0.7082 | 0.7589 | 0.7026 | 0.6204 | 0.7509 | 0.7562 | 0.7289 | 0.6824 | 0.7204 | 0.7143 | 0.6615 | 0.6932 |
| 30 | 0.7364 | 0.7521 | 0.7282 | 0.7185 | 0.7798 | 0.7730 | 0.7272 | 0.7540 | 0.6894 | 0.7424 | 0.7263 | 0.7095 |
| 36 | 0.7056 | 0.7399 | 0.7051 | 0.6779 | 0.7564 | 0.7593 | 0.7149 | 0.7010 | 0.6639 | 0.7037 | 0.6689 | 0.6709 |

The reliance signal lives in the deep layers. Averaged across steps, P1 AUC climbs from 0.5443 at layer 1 to a peak of 0.7364 at layer 30, with layer 24 close behind at 0.7082, while the shallow layers 1, 7 and 13 sit near chance throughout. Layer 30 has the highest mean AUC across steps and is fixed as the common

layer for the cross-step comparisons. The concentration of a decision-relevant feature in mid-to-late layers is consistent with the linear-representation picture of model internals [@marks2024geometry] [@zou2023repe].

## The internal signal exceeds an answer readout but fails an item-base-rate control

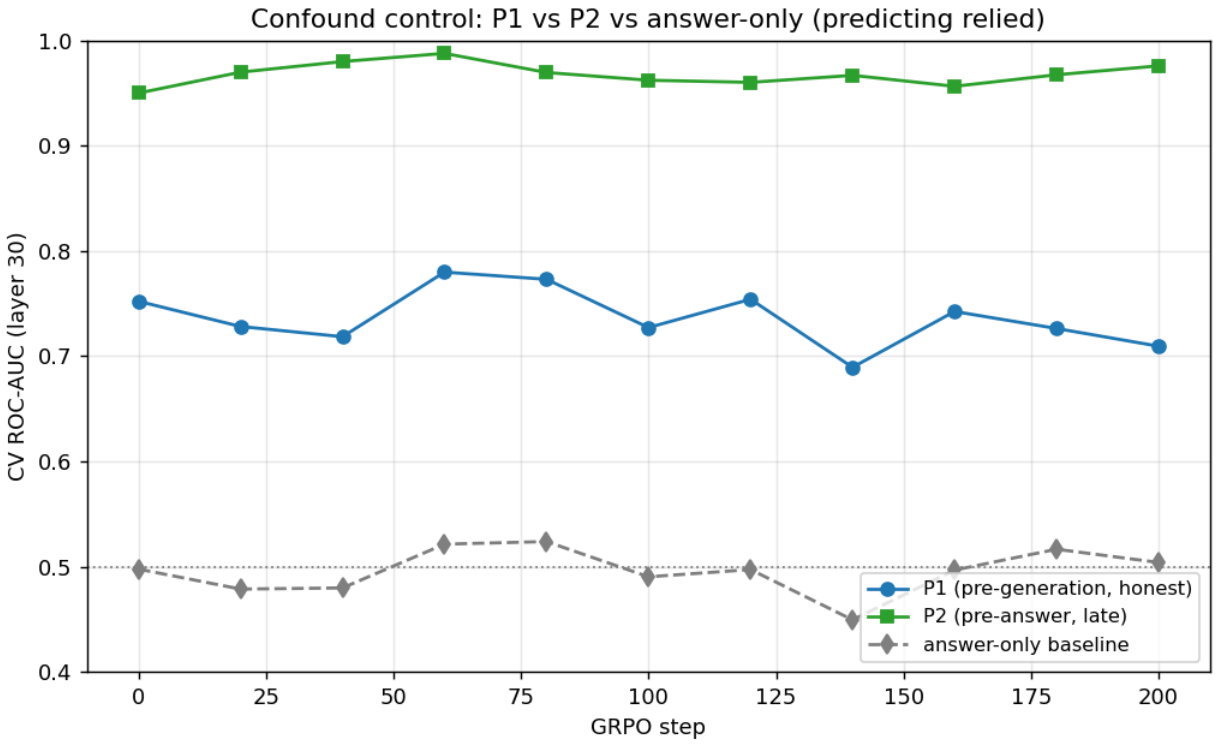


Figure 3: P1 versus P2 probe AUC against the answer-only baseline across training steps

| step | p1_auc | p2_auc | answeronly_auc |
|---|---|---|---|
| 0 | 0.7521 | 0.9499 | 0.4977 |
| 20 | 0.7282 | 0.9697 | 0.4787 |
| 40 | 0.7185 | 0.9798 | 0.4798 |
| 60 | 0.7798 | 0.9876 | 0.5214 |
| 80 | 0.7730 | 0.9694 | 0.5239 |
| 100 | 0.7272 | 0.9621 | 0.4904 |
| 120 | 0.7540 | 0.9601 | 0.4973 |
| 140 | 0.6894 | 0.9667 | 0.4495 |
| 160 | 0.7424 | 0.9563 | 0.4967 |
| 180 | 0.7263 | 0.9673 | 0.5166 |
| 200 | 0.7095 | 0.9757 | 0.5040 |

Because reliance is defined by the emitted letter, the first worry is that the probe just reads the answer. It does not: the answer-only baseline, a classifier on the one-hot emitted letter, sits at chance at every step (AUC 0.4495 to 0.5239, mean near 0.50) because the cued wrong letter varies from item to item, and the P1 probe read from the last prompt token before any generation exceeds it by a mean of 0.2404 AUC while the pre-answer P2 probe reaches 0.9499 to 0.9876. But ruling out the emitted-letter shortcut does not establish a pre-commitment representation, because a stronger confound remains: some items simply invite reliance,

a hard question with a salient cue, and the P1 probe reads the last prompt token where that item property is already fully present. We test it directly with an item-base-rate baseline, the leave-this-step-out mean reliance of each item across the other ten checkpoints, used as a predictor of per-item reliance.

| step | p1_auc | itembaserate_auc | diff | diff_ci_lo | diff_ci_hi |
|---|---|---|---|---|---|
| 0 | 0.7521 | 0.8861 | -0.1340 | -0.1909 | -0.0748 |
| 20 | 0.7282 | 0.8983 | -0.1701 | -0.2262 | -0.1124 |
| 40 | 0.7185 | 0.9061 | -0.1876 | -0.2499 | -0.1338 |
| 60 | 0.7798 | 0.9075 | -0.1277 | -0.1763 | -0.0745 |
| 80 | 0.7730 | 0.9079 | -0.1349 | -0.1883 | -0.0843 |
| 100 | 0.7272 | 0.9312 | -0.2041 | -0.2582 | -0.1477 |
| 120 | 0.7540 | 0.9637 | -0.2097 | -0.2564 | -0.1642 |
| 140 | 0.6894 | 0.9671 | -0.2777 | -0.3281 | -0.2269 |
| 160 | 0.7424 | 0.9652 | -0.2228 | -0.2733 | -0.1749 |
| 180 | 0.7263 | 0.9590 | -0.2327 | -0.2818 | -0.1848 |
| 200 | 0.7095 | 0.9478 | -0.2384 | -0.2877 | -0.1900 |
| pooled | 0.7438 | 0.9233 | -0.1795 | -0.1957 | -0.1644 |

The probe loses this comparison decisively. At every checkpoint the item-base-rate baseline beats the P1 probe, with a pooled AUC of 0.9233 against 0.7438 and a paired bootstrap difference of -0.1795 (95 percent interval -0.1957 to -0.1644, and the per-step interval excludes zero at every step), a per-step mean difference of -0.1945, and pooled balanced accuracy of 0.8620 against 0.6890. The gap against the baseline widens across training, from -0.1340 AUC at step 0 to -0.2384 at step 200, so reliance becomes more item-determined as the reward hack strengthens. The honest reading is that what the P1 probe decodes is largely static item proneness, which a trivial predictor captures better, rather than a within-step pre-commitment representation of the model's impending decision. The earlier answer-only ordering still rules out the emitted-letter echo, but it cannot license the stronger pre-commitment claim, and against item base rate the internal monitor does not survive [@belinkov2022probing] [@kadavath2022know].

## The reliance direction stays decodable while it rotates

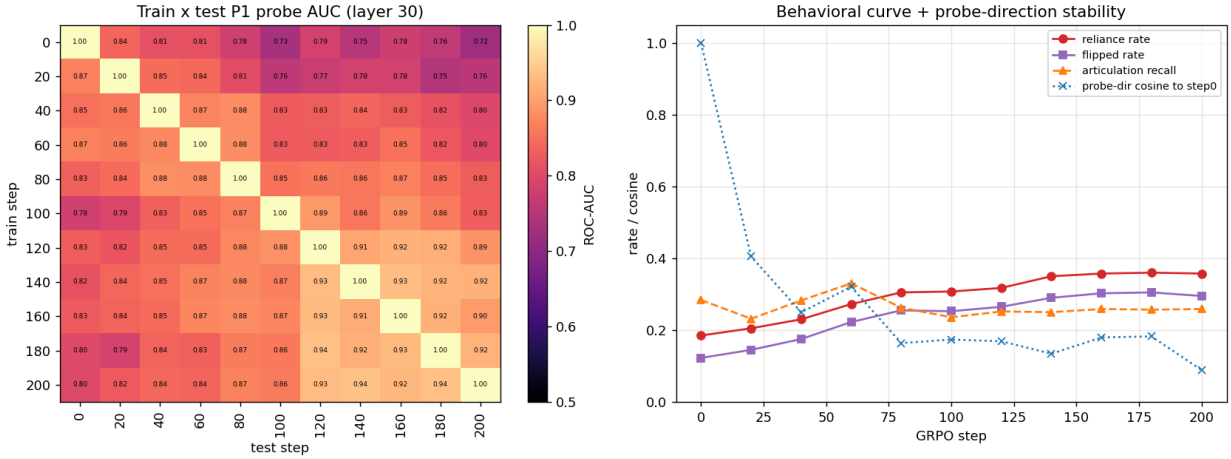


Figure 4: Cross-step probe transfer matrix with the weight-vector cosine and behavioral reliance curve

The reliance direction persists in a usable form across training even as it moves, though the honest cross-step number is lower than a same-item transfer suggests. Training a P1 probe on one checkpoint and testing on another inflates transfer when the same items appear on both sides; recomputing it item-disjoint, training

on one fixed half of the items and testing on the held-out half, gives an off-diagonal mean AUC of 0.7458, down 0.1035 from the 0.8493 obtained with item reuse (where the step 0 probe read step 200 at 0.7230), with the item-disjoint same-step diagonal averaging 0.7341, the step 0 probe reading step 200 at 0.6868 and the step 200 probe reading step 0 at 0.7482. That item-disjoint number is still well above chance, so a genuinely item-general reliance direction does persist across RL rather than being an artifact of item overlap. Yet the underlying direction rotates substantially: the cosine between adjacent checkpoints averages 0.5180 while the cosine between the step 0 and step 200 weight vectors falls to 0.0888. The model keeps an item-general, internally decodable representation of cue reliance throughout RL, but the specific direction that encodes it drifts as the behavior strengthens, so a frozen probe degrades gracefully rather than failing outright [@li2023iti].

## Limitations

The central caveat is the one the title now overstates: in this setup verbal monitorability did not decay, so we measured internal versus verbal monitoring under persistently poor verbal monitoring rather than under a verbal collapse. Whether internal monitorability survives a genuine articulation decay, the case the title asks about, requires a stronger and longer RL run that actually drives articulation toward zero, which we leave as the headline follow-up. The verbal monitor is also a deliberately simple mentions-the-cue rule rather than a strong chain-of-thought classifier, so a better verbal monitor might narrow the gap, though it would have to overcome reliance cases whose reasoning genuinely never names the cue.

The scope is narrow in other ways. The evidence is a single model, Qwen2.5-3B-Instruct, a single benchmark, ARC-Easy, and a single blatant authority cue whose presence is easy to track internally; a subtler or more natural cue might leave a fainter P1 trace. Most important, the headline internal-versus-verbal gap is an advantage over a deliberately weak lexical mentions-the-cue monitor, and the item-base-rate control in the results shows the probe itself does not beat a static item-proneness predictor, losing to it by a pooled 0.1795 AUC; a stronger LLM-judge verbal monitor and a no-cue-prompt activation probe, both deferred to follow-ups, are needed before the internal signal can be read as a model-state monitor rather than an item-property readout. The internal monitor is also one family of single-token linear probes at one layer, and the answer-only control bounds only the emitted-letter confound, not every possible shortcut. Finally, the probe is correlational: the transfer and rotation results show the direction is real and decodable, but only a causal steering test along the reliance direction would establish that the internal signal drives the behavior rather than merely accompanying it.

## Conclusion

Across a GRPO faithfulness-decay curve where cue reliance roughly doubles, an internal linear probe on pre-generation activations reads impending cue reliance with AUC between 0.71 and 0.78 at every checkpoint and beats a chain-of-thought verbal monitor at every checkpoint by a stable balanced-accuracy margin of +0.1099 (interval +0.0903 to +0.1296). That advantage over the lexical verbal monitor is real, but two controls bound what it means: the hypothesised verbal decay did not occur, so the comparison is against a persistently poor rather than a collapsing verbal monitor, and an item-base-rate baseline beats the probe itself, so the probe's score is largely a readout of static item proneness rather than a pre-commitment representation of the model's decision. What survives cleanly is a modest, item-disjoint reliance direction that persists across RL above chance at 0.7458 AUC. The practical lesson for oversight is therefore cautious: an activation monitor can be the more reliable place to look when self-report is weak, but it has to clear an item-level base-rate control before its score counts as evidence of an internal monitor [@park2024deception] [@korbak2025cotmonitorability].