# Backtracking Is Decorative: A Mechanistic Account

**Unai Zulaika** [1] **Miguel Fernandez-de-Retana** [1] **Rubén Sánchez-Corcuera** [1] **Aritz Bilbao-Jayo** [1] **Aitor Almeida** [1]

## Abstract

Reasoning language models trained with reinforcement learning frequently interrupt themselves mid-thought to "Wait, ..." or "Actually, ..." and try a different approach. It is widely assumed that this *backtracking* behavior is part of how RL post-training improves reasoning. We find it is not. The pivot decision is made *before* the model writes the interrupting token, and mechanically removing the internal signal that drives it cuts the rate of backtracking roughly in half *without changing how often the model gets the right answer* across several model families and on both math and non-math reasoning. Opening the box on a single model, we map the full pivot circuit. Three of its four parts: a sensor, an evaluator of trajectory correctness, and the downstream actuator that emits the pivot token, are already present in the base model. RL contributes the *gate* that combines the sensor and the evaluator, and even that gate is diffuse: only high-rank residual-stream interventions recover it; low-rank steering does not. How the sensor and evaluator combine depends on the domain in a predictable way: a pre-registered framework correctly predicts the gate's sign in every domain we tested. These results revise the dominant view that the most visible RL-installed reasoning behaviors are the channels through which RL improves accuracy. Decomposing what looks like a single RL-installed behavior into separable, mostly pretraining-native components opens a path toward eliciting reasoning capabilities through targeted manipulation of pre-existing structure, rather than through full RL post-training.[1]

[1]Faculty of Engineering, Universidad de Deusto, Bilbao, Spain. Correspondence to: Unai Zulaika <unai.zulaika@deusto.es>.

*Mechanistic Interpretability Workshop at the $43^{rd}$ International Conference on Machine Learning*, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

[1]Code: https://github.com/unai-zulaika/backtracking_decorative

## 1. Introduction

Recent mechanistic work locates the visible "Wait, ..." pivot in RL-trained reasoning models as a linear direction in the residual stream that pretraining already builds; in this view RL teaches the model when to engage the direction rather than installing it from scratch (Venhoff et al., 2025; 2026; Troitskii et al., 2025; Ward et al., 2025a; Yan et al., 2025). Composite steering of five such pretraining-native directions recovers up to $91\%$ of the base-to-thinking accuracy gap (Venhoff et al., 2026), and behavioral suppression of `Wait`/`Hmm` tokens at the logit level preserves accuracy across five model families (Wang et al., 2025). Backtracking is mostly pretraining-native, and the surface tokens are not strictly required for correctness.

What remains unresolved is whether the pivot decision the model makes during un-intervened generation is causally responsible for any of RL's accuracy gain. Prior work establishes the existence of pretraining-native reasoning directions (Venhoff et al., 2026) and the behavioural irrelevance of the surface tokens (Wang et al., 2025), but does not isolate one residual-stream component against accuracy under controlled intervention.

Across three model families spanning a $4\times$ scale range (Qwen3-8B, Qwen3-32B, OLMo-3-7B-Think), we find that the pivot decision is settled one token before the paragraph break by a pretraining-native direction in the residual stream, and that mechanically projecting that direction out during generation cuts the rate of backtracking by roughly half *without changing* how often the model gets the right answer. Per-trace pivot rate is uncorrelated with final-answer correctness on both math (MATH-500) and non-math reasoning (MMLU-Pro); the model that pivots less (OLMo) scores higher than the one that pivots more (Qwen3-RL) on the same problems. Three converging lines of evidence say the same thing: the surface signature of post-trained reasoning is not the channel through which RL improves accuracy.

Opening the box at Qwen3-8B (Figure 1) resolves the pivot decision into a four-part circuit: a *sensor* for rhetorical hesitation at layer 22 ($L_{22}$), an *evaluator* of trajectory correctness at $L_{27}$ near-orthogonal to the sensor, a distributed *actuator* across $L_{28}$–$L_{35}$ that emits the pivot, and a *trigger* at $L_{27}$ that combines sensor and evaluator via opposite-signed linear summation. Three components precede RL;

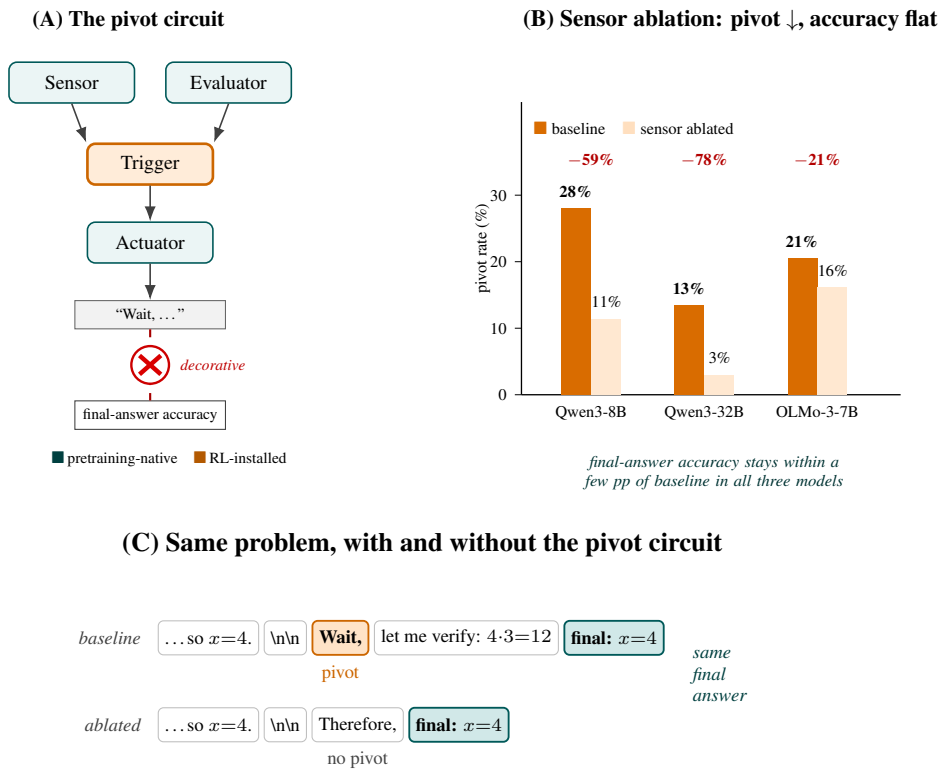


*Figure 1.* **(A)** Pivot circuit: three components are pretraining-native (teal); RL adds only the trigger (orange). The broken arrow marks the absence of an accuracy effect. **(B)** Projecting the sensor out cuts pivot rate by 21–78% across three model families; accuracy stays flat. **(C)** Same problem with the circuit intact (top, pivot emitted) and with the sensor projected out (bottom, pivot skipped). Both reach the same answer.

only the trigger is RL-installed, and even the trigger is not a compact gate: a six-rung weight-localization ladder is null at narrow scope, and the smallest succeeding subset spans the embedding together with the first 28 transformer layers. The gate's sign flips by domain in a predictable way (rescue vs. revision), with a pre-registered framework correctly predicting the gate's sign in every MMLU-Pro domain we tested.

We make four claims, established on Qwen3-8B-Base/RL with the universal claims replicated across families and scales: **(1)** a cross-family pre-emission pivot sensor, decodable in the residual stream exactly one token before each paragraph break with a discrete cliff at that position, replicating on out-of-corpus AIME 2022–25 and across a $4\times$ scale range (§4); **(2)** mechanism-level isolation of the decorative pivot, extending the behavioural suppression result of Wang et al. (2025) to the residual-stream level, with accuracy preserved across ablation, observation, and cross-model contrast on math and non-math reasoning (§5); **(3)** a four-part pivot circuit (sensor, evaluator, actuator, trigger) of which only the trigger is RL-installed, and it is high-rank linear in the activation difference (six narrow weight-localization attempts null) and weight-spread across the embedding and the first 28 transformer layers, hinting that RL's contribution may not localize at the granularities most interpretability tools currently target (§6); **(4)** an opposite-signed linear gate whose sign flips between rescue and revision domains in a way our pre-registered framework correctly predicts in every MMLU-Pro domain we tested (§6).

## 2. Related Work

**Backtracking as a residual-stream direction.** A growing line of work shows that backtracking in post-trained reasoning models is carried by one or a small number of linear residual-stream directions (Venhoff et al., 2025; 2026; Yan et al., 2025; Ward et al., 2025a), with Venhoff et al. (2026) framing pretraining as building the reasoning machinery and RL as learning when to engage it. The closest precedent to our pre-emission analysis is Troitskii et al. (2025), who probe the residual stream immediately before each `wait` token via sparse crosscoders and identify features that promote and suppress wait emission. We extend this line by giving a causal account of the full pivot circuit (sensor, evaluator, trigger, actuator) and by isolating each component via projection-out interventions during generation. Token- and feature-level analyses of pivot tokens (Qian et al., 2025;

Cox et al., 2026; Hwang et al., 2026; Galichin et al., 2026) complement this residual-stream account at adjacent granularities. The decorative reading we develop is also continuous with the broader chain-of-thought-faithfulness literature (Lanham et al., 2023; Turpin et al., 2023), which establishes that surface CoT tokens can systematically misrepresent the model's actual reasoning process. The behavioral analogue of our decorative-for-accuracy claim is Wang et al. (2025), who suppress `Wait`/`Hmm` tokens at the logit level and report unchanged accuracy across five model families; we extend that result to the mechanism level by isolating and removing the upstream signal that drives the behavior. Methodologically, our pre-emission probe is closest to probe-gated activation steering (Cheng et al., 2025).

**What RL post-training adds, and circuit motifs.** A complementary line shows that RL's contribution to reasoning capability is sparse and largely preserves base-model representations (Sinii et al., 2025b;a; Yue et al., 2025; Shi et al., 2026; Zhao et al., 2025; Meng et al., 2026; Ward et al., 2025b); Meng et al. (2026) report that more than 83% of token positions have near-zero $\mathrm{KL}(\pi_{\mathrm{RL}} \,\|\, \pi_{\mathrm{base}})$ under DAPO; Huang et al. (2026) measure RL direction by per-token $\log \pi_{\mathrm{RL}} - \log \pi_{\mathrm{base}}$ on emitted tokens and find the signal concentrated on roughly 7–11% of rare positions; and Ward et al. (2025b) show that a rank-1 LoRA can recover most of the base-to-thinking accuracy gap; and Zhu et al. (2025) argue that RLVR updates concentrate *off* the principal weight directions. The load-bearing RL modification we identify (the trigger) is similarly diffuse at the parameter-support level, and our isolation result adds that one specific component of that sparse contribution (backtracking) is causally inert for accuracy. Anthropic's attribution-graph circuit tracing on Claude 3.5 Haiku (Ameisen et al., 2025; Lindsey et al., 2025) reveals multi-step reasoning circuits at the feature-graph level; our four-part decomposition is at a coarser, behavior-specific granularity and complements that line by isolating where RL's modification lives within an existing pretraining-native pipeline. This suggests the bulk of the RL gain lies in the other base-model directions steered by prior composite-recovery protocols (Venhoff et al., 2026). The opposing-sign motif itself is precedented in pretrained transformers, both in the IOI S-Inhibition circuit (Wang et al., 2023) and in copy-suppression heads (McDougall et al., 2023); less precedented is an RL-installed opposing-sign gate over pretraining-native features at *different* layers, with the combination approximately linear rather than logical. The linear-representation hypothesis (Park et al., 2024) frames the sensor and evaluator directions we identify as empirical instances. Appendix F expands this comparison.

## 3. Setup

**Models.** The headline analysis is on Qwen3-8B-Base and Qwen3-8B-RL (Yang et al., 2025), an outcome-based-RL reasoning model trained on mathematics (`Qwen/Qwen3-8B-Base` and `Qwen/Qwen3-8B`). Cross-family replication uses OLMo-3-7B-Think (Team OLMo et al., 2025) (`allenai/Olmo-3-7B-Think`, Llama-derived, 32 layers, $d_{\text{model}}$=4096). Cross-scale replication uses Qwen3-32B (`Qwen/Qwen3-32B`, 64 layers, $\sim$4$\times$ parameter count of the 8B). All three families are post-trained reasoning models that emit thinking-style chains of thought.

**Datasets.** The headline probe is fit on a trace bank covering MATH-500 (Lightman et al., 2024; Hendrycks et al., 2021) reasoning chains and the Venhoff-500 reasoning corpus (Venhoff et al., 2025), totaling 1,445 trace groups and 13,299 break-labeled residuals at $(L_{22}, p{-}1)$. The centerpiece compositional ablation (§6) is run at $N$=100 MATH-500 problems per condition, aggregated across four seeds. Cross-domain replication uses MMLU-Pro (Wang et al., 2024) non-math (13 categories, of which 6 yield usable per-category correct/incorrect class balance for evaluator extraction). Out-of-corpus contamination defense uses AIME 2022, 2023, 2024, and 2025 (Appendix C, $N$=119 problems with 4,040 break labels). Cross-family OLMo runs use $N$=50 at two seeds and $N$=100 at one seed; cross-scale Qwen3-32B runs use $N$=25 at two seeds (Appendix D).

**Paragraph breaks and labels.** A paragraph break is any token position $p$ in a generated trace matching the regular expression `[.?!]\n(\n?)` immediately following a content token; we write $p{-}1$ for the last content token before the break, the pre-emission position our probe targets. We label every break into {pivot, continuation, enumeration} by the first content token of the next paragraph using a deterministic classifier (Appendix G). Pivot openers include `Wait`, `Alternatively`, `Hmm`, `Actually`; enumeration openers match numeric or bullet markers (`1.`, `2)`, `(a)`, `-`, `*`); all other openers count as continuation. The OLMo classifier uses an OLMo-native pivot lexicon derived from the empirical top-first-words after breaks in OLMo's own generations.

**Residual extraction.** For every labeled break at token position $p$ we cache residual-stream activations at the surrounding token positions $p{-}k$ and $p{+}k$ for $k$=1, . . . , 10, at the probed layers. At Qwen3-8B we use layers $\{10, 22, 30\}$ for the layer-comparison probe and the full stack $\{0, \ldots, 35\}$ for the activation-patch sweep (§6). Each cache entry is grouped by the trace it came from, so that the leave-one-trace-out cross-validation below places all residuals from a single trace into the same fold; the probe is never trained on

residuals drawn from the same generation it is evaluated on.

**Probe.** The pre-emission probe is a logistic classifier on cached residuals: StandardScaler → PCA-128 → logistic regression ($C$=1, liblinear solver). We report AUC under leave-one-trace-out (LOPO) cross-validation over the 1,445 trace groups, with 95% confidence intervals computed by $B$=1000 cluster-bootstrap resamples over groups. Hyperparameter robustness is in Appendix A.

**Sensor and evaluator directions.** The sensor direction $\mathbf{v}_{\mathrm{BT}}$ is the unit-normed centroid difference between pivot and continuation residuals at $(L_{22}, p{-}1)$ in Qwen3-8B, and at the analogous probe-peak layer in each cross-family / cross-scale model:

$$\mathbf{v}_{\mathrm{BT}} \;=\; \frac{\mu_{\mathrm{piv}} - \mu_{\mathrm{cont}}}{\|\mu_{\mathrm{piv}} - \mu_{\mathrm{cont}}\|_2}. \tag{1}$$

The evaluator direction $\mathbf{w}_E$ is defined by the same recipe applied to final-correct vs final-incorrect trajectory residuals at $(L_{27}, p{-}1)$ in Qwen3-8B. In Qwen3-32B we identify the clean evaluator location at $L_{62}$ via a within-layer cosine sweep $\cos(\mathbf{v}_{\mathrm{BT}}^{(\ell)}, \mathbf{w}_E^{(\ell)})$, avoiding the sensor-contaminated linear-probe peak (Appendix D). Both directions are used for steering and for projection-out interventions during generation. $\mathbf{v}_{\mathrm{BT}}$ and $\mathbf{w}_E$ correspond to the *sensor* and *evaluator* of the four-part pivot circuit (sensor, evaluator, trigger, actuator) we map causally in §6; the remaining two components are introduced there.

**Generation and ablation protocol.** For ablation experiments we sample with $T$=0.7, top-$p$=0.95, max-new-tokens 4096, varying seed $\in \{0, 1, 2, 3\}$ where indicated. Sensor- and evaluator-projection-out interventions are forward-pre-hooks at the relevant layer that replace the last-position residual $h$ with its component orthogonal to the named direction at every generation step:

$$h' \;=\; h - (\hat{v}^{\top} h)\, \hat{v}, \tag{2}$$

where $\hat{v}$ is $\mathbf{v}_{\mathrm{BT}}$ or $\mathbf{w}_E$ as appropriate, and $h'$ is the residual that the rest of the forward pass sees.

**Metrics and notation.** $L_\ell$ refers to the residual stream after decoder layer $\ell$. The *pivot rate* of a generation run is the fraction of paragraph breaks classified as pivots by the deterministic classifier above; *accuracy* is the fraction of problems whose final answer matches the dataset ground truth. We say a behavior is *decorative for accuracy* if mechanically removing the residual-stream signal that drives it produces a large change in the rate of the behavior together with no detectable change in accuracy under the same generation protocol. Effect sizes are reported as deltas in percentage points (pp).

## 4. A Universal Pre-Emission Sensor for Pivots

We establish the sensor in three steps. First, a logistic probe on Qwen3-8B-Base residuals separates pivots from continuations at AUC 0.908 at the last pre-break token, with a discrete cliff at $p{-}1$ that does not appear at early layers (Figure 2a). Second, the same direction extracted by centroid difference is causally a positive driver of pivot emission: projecting it out during generation cuts the pivot rate by roughly half (Table 1). Third, both the probe peak and the causal effect replicate on out-of-corpus contest problems and on two additional reasoning-model families spanning a $4\times$ scale range (Figure 2b).

**The pivot decision is decodable one token before the break.** A logistic probe at $(L_{22}, p{-}1)$ on Qwen3-8B reaches AUC $\mathbf{0.908}$ (95% CI $[0.902, 0.914]$, leave-one-trace-out over 1,445 groups), and transfers between Base and RL generations within noise. Full per-layer-by-offset sweep in Appendix A.

**The pre-emission state is a discrete cliff, not a distance ramp.** Two effects in Figure 2a are jointly informative. First, at $L_{22}$ and $L_{30}$ the AUC at $p{-}1$ is separated from the AUC at $p{-}2$ by a step whose CIs do not overlap (a 2.6 pp drop at $L_{22}$, a 4.5 pp drop at $L_{30}$); at $L_{10}$ the same comparison is within noise and the curve is a smooth distance-to-break ramp. Second, beyond $k$=2 the AUC decays slowly with distance at every layer. The combined picture is a depth–time ladder: content-adjacent features correlate gradually with proximity to a break at early layers, while the pivot decision itself becomes a discrete state at the last pre-break token in mid-late layers.

**Removing the direction during generation cuts pivot rate by roughly half.** The unit-normed centroid-difference direction $\mathbf{v}_{\mathrm{BT}}$ extracted at $(L_{22}, p{-}1)$ via Eq. 1 is causally a positive driver of pivot emission. Projecting $\mathbf{v}_{\mathrm{BT}}$ out of the last-position $L_{22}$ residual at every generation step via Eq. 2 drops Qwen3-8B-RL's pivot rate from 28.0% baseline to 11.4% ($\Delta = -16.6$ pp; mean across $N$=100 MATH-500 problems and four seeds). A random-unit-vector projection control at the same layer is flat at baseline; final-answer accuracy is preserved in every condition, the entry point for §5.

**The result is not an artifact of MATH-500 contamination.** To rule out that the headline AUC of 0.908 is inflated by Qwen3-8B having seen MATH-500 during pretraining, we repeat the probe on Qwen3-8B-RL traces of AIME 2022, 2023, 2024, and 2025 problems ($N$=119 across years, 4,040 break labels, 117 trace groups). The combined-corpus $L_{22}$ AUC is $\mathbf{0.903}$, within 0.5 pp of the MATH-500 headline; the post-cutoff AIME 2025 subset alone gives $L_{22}$

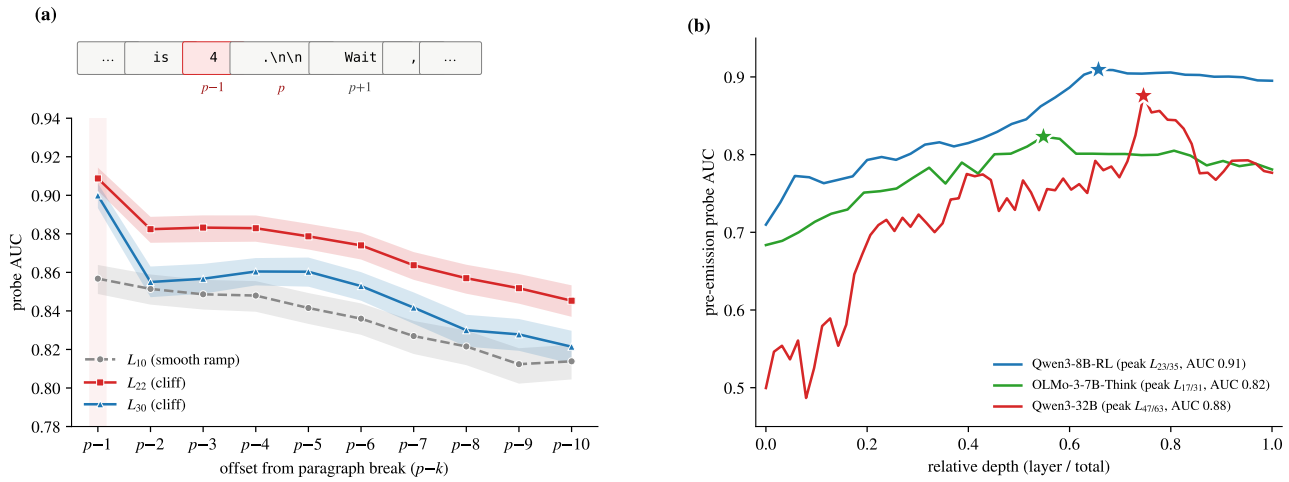


*Figure 2.* **Pre-emission probe is sharply localized and replicates across model families.** *(a)* Probe AUC vs. pre-break offset on Qwen3-8B-Base with 95% cluster-bootstrap CIs. The token strip above shows the position convention: `4` at $p-1$, `.\n\n` at $p$. At $L_{22}$ and $L_{30}$ the AUC at $p-1$ does not overlap the AUC at $p-2$; at $L_{10}$ it does. *(b)* Per-layer probe AUC for three model families. Stars mark per-model peaks; all three fall in the mid-to-late stack.

AUC 0.847 (Appendix C). The pre-emission cliff replicates on out-of-corpus contest problems.

**The same sensor exists in OLMo-3-7B-Think and Qwen3-32B at the same relative depth.** Both the probe peak and the causal effect replicate on OLMo-3-7B-Think and Qwen3-32B (Figure 2b, Table 1). All three peaks fall in the mid-to-late stack (0.55 to 0.75 relative depth, AUC within 9 pp of the Qwen3-8B headline). Sensor projection-out drops pivot rate by $-4.4$ pp on OLMo (three runs) and $-10.4$ pp on Qwen3-32B (two seeds), with accuracy preserved and random-vector controls flat. Full per-layer profiles in Appendices E and D.

## 5. Pivots Are Decorative for Accuracy

The decorative claim follows from three converging lines of evidence: under our causal interventions, pivot rate drops by 21–78% but accuracy is preserved (Table 1); per-trace pivot rate is uncorrelated with accuracy on both math and non-math reasoning (Figure 3); and the cross-model contrast (OLMo pivots less than Qwen3-RL but scores higher) is the same pattern at the family level.

**Ablation preserves accuracy across all models and seeds.** Across every condition $\times$ seed cell of our compositional ablation, final-answer accuracy stays within $\pm 2$ pp of baseline while pivot rate drops by 21–78% (Table 1). The intervention crashes the surface signature without moving the metric the model is being evaluated on.

**Per-trace correlation is null on both math and non-math reasoning.** The same null obtains in pure observation: per-problem pivot rate is uncorrelated with final-answer correctness on MATH-500 and on MMLU-Pro non-math (Figure 3; both Pearson 95% CIs include zero). The pivot behavior the sensor produces is statistically independent of whether the model gets the right answer, in the un-intervened generation as well as under causal ablation.

**Cross-model contrast: less pivoting, more accuracy.** The same dissociation holds at the model-family level: OLMo-3-7B-Think pivots roughly $1.5\times$ less often than Qwen3-8B-RL on MATH-500 yet scores $\sim 22$ pp higher on accuracy. The model that pivots less, scores more.

**Decorative in isolation, not necessarily in composite.** Venhoff et al. (2026) recover up to 91% of the base-to-thinking accuracy gap by activating five base-model directions in concert (verification, hypothesis-checking, knowledge-retrieval, uncertainty-management, backtracking). Our result is consistent and adds a controlled-isolation finding: backtracking is causally inert for accuracy when isolated. We rule out that backtracking carries the accuracy gap by itself; by elimination, the accuracy-bearing capacity sits primarily in the other four arms. What we cannot rule out is that backtracking contributes to accuracy only in interaction with those directions, an interaction the isolated-ablation protocol cannot expose.

## 6. Opening the Box: A Pivot Circuit at Qwen3-8B

Having established a universal sensor and shown that the pivots it drives are decorative for accuracy, we now open the box on one model, Qwen3-8B-RL, and ask how the pivot

*Table 1.* **Sensor projection-out drops pivot rate but not accuracy across three model families.** Arrows are baseline → ablated. Pivot rate drops in every model; accuracy stays within ±2 pp of baseline. Random-vector controls are flat (not shown).

| Model | Layer | $N\times$seeds | Pivot rate | $\Delta$ | Accuracy |
|---|---|---|---|---|---|
| Qwen3-8B-RL | $L_{22/35}$ | 100×4 | 28.0% → 11.4% | −16.6 pp | 80% → 83% |
| OLMo-3-7B-Think | $L_{17/31}$ | 50×3 | 20.5% → 16.1% | −4.4 pp | 82% → 84% |
| Qwen3-32B | $L_{47/63}$ | 25×2 | 13.4% → 3.0% | −10.4 pp | 80% → 74% |

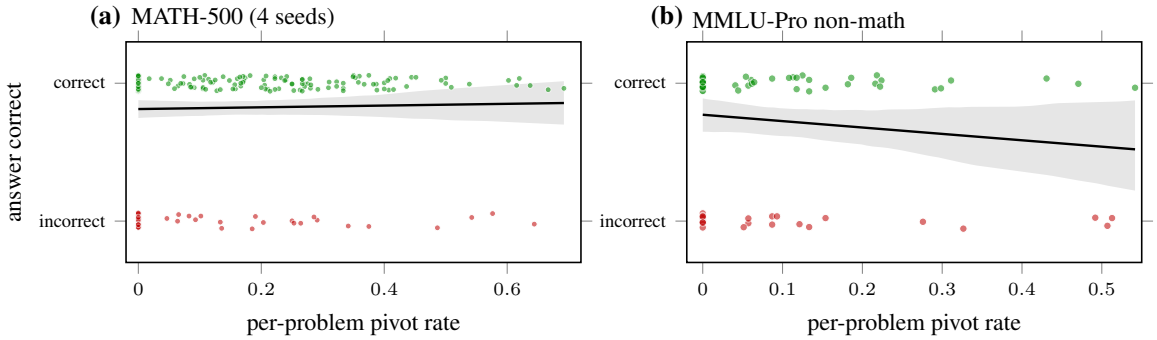


*Figure 3.* **Per-problem pivot rate is uncorrelated with final-answer accuracy on math and non-math reasoning.** Each dot is one baseline trace from Qwen3-8B-RL; green dots are correct, red are incorrect. *(a)* MATH-500, 4 seeds, $N$=225, Pearson $r$=+0.029, 95% CI [−0.11, +0.15]. *(b)* MMLU-Pro non-math, $N$=80, $r$=−0.148, 95% CI [−0.38, +0.10]. Black line: linear fit; grey band: 95% bootstrap CI. Both CIs cross zero. Y-jittered for visibility.

circuit is wired. Four components emerge: a *sensor* for rhetorical hesitation at $L_{22}$ (the direction $\mathbf{v}_{\mathrm{BT}}$ defined in §3), an *evaluator* for trajectory correctness one layer downstream at $L_{27}$ (the direction $\mathbf{w}_E$), a *trigger* that reads $\mathbf{v}_{\mathrm{BT}}$ and $\mathbf{w}_E$ at $L_{27}$ and writes a residual state the actuator can consume, and a distributed *actuator* across $L_{28}$–$L_{35}$ that converts that state into the next-token pivot logit. We use *trigger* for the RL-installed component and *compositional gate* for the linear-summation operation that the trigger implements over $\mathbf{v}_{\mathrm{BT}}$ and $\mathbf{w}_E$. Three of the four components are pretraining-native (sensor, evaluator, actuator). RL contributes only the trigger, and the trigger is high-rank linear in the activation difference at $L_{27}$, diffuse across the embedding and the first 28 transformer layers in weight-space (Figure 5).

**The sensor is orthogonal to content and to enumeration.** Projecting $\mathbf{v}_{\mathrm{BT}}$ out of the $L_{22}$ residual at every generation step preserves MATH-500 accuracy on a paired-rollout test, so the pivot direction is linearly separable from the content the model uses to solve the problem. It is also orthogonal to a second structural mode of breaking a paragraph: a three-way probe over {pivot, continuation, enumeration} at $(L_{22}, p{-}1)$ separates all three classes at AUC $\geq 0.92$ pairwise, and the centroid-difference direction for enumeration has cosine $+0.04$ with $\mathbf{v}_{\mathrm{BT}}$. Two distinct structural modes occupy independent linear subspaces at the pre-emission position.

**What $\mathbf{v}_{\mathrm{BT}}$ encodes lexically.** The direction tracks rhetorical hesitation rather than a forward model of the trajectory's failure. Logit-lens at Base $(L_{22}, p{-}1)$ ranks pivot-opener lexemes high (`anyways`, `anyway`, `Maybe`, `maybe`, `Anyway`); the same words are absent from the continuation top-20. The max-activating contexts of $\mathbf{v}_{\mathrm{BT}} \cdot h_{L_{22}}$ are dominated by clause-terminal punctuation immediately before pivot openers (e.g. `?` before `Wait, let me reconsider`).

**A near-orthogonal pretraining-native evaluator at $L_{27}$.** A logistic probe for trajectory correctness at $(L_{27}, p{-}1)$ reaches AUC 0.83 in both Base and RL (gap $+0.009$); the wrongness signal is not RL-installed. We extract $\mathbf{w}_E$ as the centroid-difference between correct and incorrect residuals at this position. Within-layer $\cos(\mathbf{v}_{\mathrm{BT}}, \mathbf{w}_E) = -0.004$: sensor and evaluator are essentially orthogonal.

**The compositional gate is opposite-signed and approximately linear.** How do the two signals combine during generation? Ablating each independently and jointly on Qwen3-8B-RL gives the result in Table 2. $\mathbf{v}_{\mathrm{BT}}$ is a positive driver (ablation drops pivot rate by $-16.6$ pp). $\mathbf{w}_E$ enters with the *opposite* sign (ablation *raises* pivot rate by $+2.5$ pp). Their joint ablation is sub-additive against the linear-additive prediction by $-3.0$ pp, consistently across seeds. The two signals therefore combine approximately linearly with opposite signs, not through a non-linear AND-NOT gate. Final-answer accuracy stays within ±2 pp of baseline in every cell.

*Table 2.* **Compositional ablation of $\mathbf{v}_{\mathrm{BT}}$ and $\mathbf{w}_E$ on Qwen3-8B-RL.** Pivot rate is the mean over $N$=100 MATH-500 problems and four seeds; signs are consistent across seeds. Joint ablation is sub-additive vs. the linear-additive prediction by $-3.0$ pp on average.

| condition | pivot rate | $\Delta$ vs baseline | accuracy |
|---|---|---|---|
| baseline | 28.0% | — | 80% |
| ablate $\mathbf{v}_{\mathrm{BT}}$ at $L_{22}$ | 11.4% | $-16.6$ pp | 83% |
| ablate $\mathbf{w}_E$ at $L_{27}$ | 30.5% | $+2.5$ pp | 79% |
| ablate both | 11.0% | $-17.0$ pp | 81% |

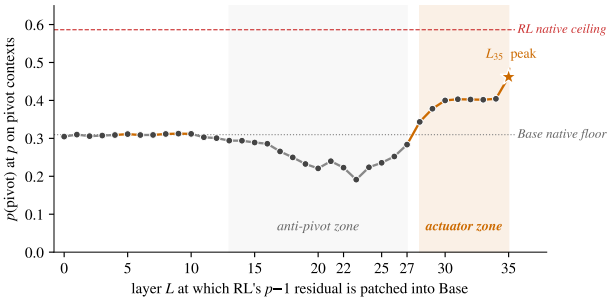


*Figure 4.* **The actuator is localized to $L_{28}-L_{35}$.** For each layer $L$, the curve shows $p(\mathrm{pivot})$ at the next-token position when RL's $p-1$ residual is patched into Base at $L$. Patching at $L_{13}-L_{27}$ suppresses pivots below the Base floor (anti-pivot zone, gray); patching at $L_{28}-L_{35}$ promotes them toward the RL ceiling (actuator zone, orange), peaking at $L_{35}$.

**The actuator is localized to the late stack at $L_{28}-L_{35}$.** Where in the network does the gated state actually convert into a pivot token? We sweep the activation patch one layer at a time: at every layer $L$, we copy the $p-1$ residual from RL into Base, let Base's forward pass continue from there, and measure $p(\mathrm{pivot})$ at the next-token position on pivot contexts (Figure 4). The result resolves the actuator into two zones. Patching at $L_{13}-L_{27}$ *suppresses* pivots below Base's native floor (an anti-pivot zone). Patching at $L_{28}-L_{35}$ *promotes* pivots, ramping toward RL's ceiling and peaking at $L_{35}$ at $+15.2$ pp above the Base floor. The actuator is therefore localized to the last quarter of the stack, attention-mediated, and present in Base: when the upstream residual at $L_{27}$ is set appropriately, the late layers pick it up and emit the pivot token without any further RL-installed state.

**The trigger is high-rank, diffuse, and not localized to any narrow weight support.** Patching the full $L_{27}$ residual from RL into Base during generation recovers $\sim 27\%$ pivot rate (matching RL's native rate, against Base's $\sim 11\%$ floor). Holding that activation-patch ceiling fixed, we walk a localization ladder of progressively wider interventions (Figure 5). Six narrow interventions are null at the Base floor: rank-$K$ direction injection at $L_{27}$ for $K \leq 128$, single attention head $L_{35}H_{12}$, $L_{27}$ sublayer patches (attention-only, MLP-only, or both), single-layer $L_{27}$ weight transplant, and $L_{22}-L_{27}$ band weight transplant. Only wider interventions recover any pivot mass: a rank-256 direction at $L_{27}$ reaches the Base floor, and transplanting the embedding together with all weights of layers $L_0-L_{27}$ (309 parameter tensors) gives $\sim 60\%$ of the activation-patch ceiling. RL did not install a compact gate at a narrow locus; it made small adjustments across the embedding and the first 28 layers, and the residual state those adjustments produce at $L_{27}$ output is what the actuator reads.

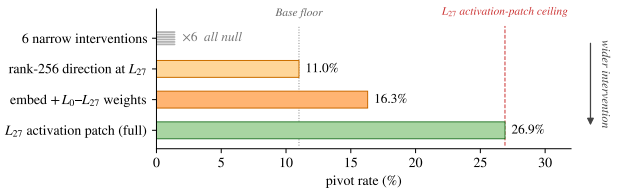


*Figure 5.* **The trigger has no narrow weight support.** Each row is one intervention during Qwen3-8B generation. Six narrow attempts are null at the Base pivot floor; only wide-scope interventions recover pivot mass. The smallest succeeding weight subset is the embedding plus all weights of layers $L_0$–$L_{27}$ (309 tensors), reaching $\sim 60\%$ of the $L_{27}$ activation-patch ceiling.

**The gate's sign is set by the domain's error structure: pivot-as-rescue vs pivot-as-revision.** On math, incorrect trajectories pivot more often than correct ones (pivoting is a *rescue* signal); the centroid $\mathbf{w}_E$ therefore points away from "currently pivoting", and ablating it raises pivot rate. On law, psychology, and economics, the asymmetry inverts (pivoting is constitutive of careful *revision*, not a symptom of failure); the per-domain $\mathbf{w}_E$ points toward pivoting, and ablating it drops pivot rate. We pre-registered the predicted sign for six MMLU-Pro domains from natural-trace correctness asymmetries; every prediction matched, including economics, where the prediction was made before observing the ablation result, and signs held when replicated at higher $N$. The sign on which $\mathbf{w}_E$ enters the gate is therefore set by the domain's error-pivot association, not by the math training distribution.

## 7. Discussion and Conclusion

**Where the accuracy actually lives.** The decorative finding holds across causal isolation, per-trace observation, and cross-model contrast, and extends the behavioural result of Wang et al. (2025) to the mechanism level. The natural reading is that the bulk of the base-to-thinking accuracy gap recovered by composite steering (Venhoff et al., 2026) lives in the other four directions (verification, hypothesis-checking, knowledge-retrieval, uncertainty-management). Applying our pipeline to each is an immediate program for future work.

**The trigger is diffuse, and what that means for the linear-representation reading.** Venhoff et al. (2026) frame pretraining as building the reasoning machinery and RL as learning when to engage it; our case study refines this. The engagement is mediated by a four-part circuit, of which three components (sensor, evaluator, actuator) precede RL and only the trigger is RL-installed. Six progressively wider weight-localization attempts are null at the Base floor (Figure 5); the smallest succeeding subset is the embedding plus all weights of layers $L_0$–$L_{27}$. RL's contribution is better described as small adjustments across the entire upstream pipeline, with a residual-state effect at $L_{27}$ that the pretraining-native actuator reads, consistent with the sparse-but-distributed signal of Meng et al. (2026). Read against the Linear Representation Hypothesis (Park et al., 2024), this asymmetry is informative: the two pretraining-native directions $\mathbf{v}_{\mathrm{BT}}$ and $\mathbf{w}_E$ are clean LRH-compatible objects, but the RL-installed trigger is not, and its high-rank character parallels documented limits of LRH-compatible decomposition for compound concepts (Chanin et al., 2025). Single-direction analyses can identify the inputs to the gate but not the gate itself.

**Methodological implications.** Three implications fall out of the case study. (i) Single-feature methods can miss an opposing-sign inhibitor: rank-$K$ steering at $L_{27}$ was null while $\mathbf{w}_E$ was firing, and unlocked at $K{=}128$ once $\mathbf{w}_E$ was projected out. The opposing-sign motif is precedented in pretrained transformers (Wang et al., 2023; McDougall et al., 2023) but can be RL-installed across distinct layers as we observe. (ii) Small-$N$ pilots can mis-identify interaction terms: our $N{=}25$ pilot gave a $+12.9$ pp super-additivity reading consistent with an AND-NOT gate, which did not survive replication at $N{=}100$ across four seeds (mean $-3.0$ pp). Paired ablation at $N \gtrsim 100$ with multiple seeds is the minimal diagnostic before any super-additivity claim. (iii) Within-layer cosine sweeps are needed to find the clean evaluator at scale: the sensor-contaminated $L_{47}$ peak in Qwen3-32B gave a noisy readout while the near-orthogonal $L_{62}$ gave cleaner geometry.

**The rescue/revision framework as a predictive lens.** The sign on which the evaluator enters the gate flips by domain: rescue domains (incorrect trajectories pivot more, e.g. math) give an opposite-signed gate; revision domains (correct trajectories pivot more, e.g. law and psychology) give a same-signed gate. The natural-trace correctness asymmetry is the predictor, and the framework correctly predicts the sign in every MMLU-Pro domain we tested. The same recipe applied to other reasoning behaviors should yield domain-dependent signs, extending mechanism case studies cross-domain without re-running the full causal protocol.

**Limitations and conclusion.** The full four-part circuit is established only at Qwen3-8B-RL on math; evaluator and gate do not robustly transfer to OLMo-3-7B-Think or Qwen3-32B at the pilot $N$ available. The decorative finding tests backtracking *in isolation*: we rule out backtracking-by-itself as RL's accuracy channel, not backtracking-as-component-of-a-larger-composite. With those qualifications, the pivot tokens of post-trained reasoning are not the channel through which RL improves accuracy, and the only RL-installed component of the surrounding circuit is a high-rank, diffuse modification spanning the embedding and the first 28 layers, which single-direction interventions will systematically miss.

## Acknowledgements

This work was funded by the European Union's Horizon Europe research and innovation programme through the project IDEA4RC: *Intelligent Ecosystem to improve the governance, the sharing and the re-use of health Data for Rare Cancers* (grant agreement ID 101057048), and through the project ProtectChild: *A PRivacy-prOTecting Environment for Child Transplants health-related and genomic data integration in the European Reference Network* (grant agreement ID 101137423).

## References

Ameisen, E., Lindsey, J., Pearce, A., Gurnee, W., Turner, N. L., Chen, B., Citro, C., Abrahams, D., Carter, S., Hosmer, B., Marcus, J., Sklar, M., Templeton, A., Bricken, T., McDougall, C., Cunningham, H., Henighan, T., Jermyn, A., Jones, A., Persic, A., Qi, Z., Thompson, T. B., Zimmerman, S., Rivoire, K., Conerly, T., Olah, C., and Batson, J. Circuit tracing: Revealing computational graphs in language models. Transformer Circuits Thread. https://transformer-circuits.pub/2025/attribution-graphs/methods.html, 2025.

Chanin, D., Wilken-Smith, J., Dulka, T., Bhatnagar, H., Golechha, S., and Bloom, J. A is for Absorption: Studying feature splitting and absorption in sparse autoencoders. In *Advances in Neural Information Processing Systems*, 2025.

Cheng, Z., Gan, J., Jiang, Z., Wang, C., Yin, Y., Luo, X., Fu, Y., and Gu, Q. Steering when necessary: Flexible steering large language models with backtracking. In *Advances in Neural Information Processing Systems*, 2025.

Cox, K., Kianersi, D., and Garriga-Alonso, A. Decoding answers before chain-of-thought: Evidence from pre-CoT probes and activation steering. *arXiv preprint arXiv:2603.01437*, 2026.

Galichin, A. V., Dontsov, A., Druzhinina, P., Razzhigaev, A., Rogov, O., Tutubalina, E., and Oseledets, I. I have covered all the bases here: Interpreting reasoning features in large language models via sparse autoencoders. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 30771–30779, 2026.

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. Measuring mathematical problem solving with the MATH dataset. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2021.

Huang, K., Meng, H., Wu, J., Lu, J., Ma, C., Chen, Z., Wang, X., Ding, B., Wu, J., Wang, X., et al. On the direction of RLVR updates for LLM reasoning: Identification and exploitation. *arXiv preprint arXiv:2603.22117*, 2026.

Hwang, J., Heo, B., Yun, S., and Han, D. Oops, Wait: Discourse tokens matter in reasoning model. *arXiv preprint arXiv:2601.17421*, 2026.

Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C., Hernandez, D., Li, D., Durmus, E., Hubinger, E., Kernion, J., et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., and Cobbe, K. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2024.

Lindsey, J., Gurnee, W., Ameisen, E., Chen, B., Pearce, A., Turner, N. L., Citro, C., Abrahams, D., Carter, S., Hosmer, B., Marcus, J., Sklar, M., Templeton, A., Bricken, T., McDougall, C., Cunningham, H., Henighan, T., Jermyn, A., Jones, A., Persic, A., Qi, Z., Thompson, T. B., Zimmerman, S., Rivoire, K., Conerly, T., Olah, C., and Batson, J. On the biology of a large language model. Transformer Circuits Thread. https://transformer-circuits.pub/2025/attribution-graphs/biology.html, 2025.

McDougall, C., Conmy, A., Rushing, C., McGrath, T., and Nanda, N. Copy suppression: Comprehensively understanding an attention head. *arXiv preprint arXiv:2310.04625*, 2023.

Meng, H., Huang, K., Wei, S., Ma, C., Yang, S., Wang, X., Wang, G., Ding, B., and Zhou, J. Sparse but critical: A token-level analysis of distributional shifts in RLVR fine-tuning of LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026.

Park, K., Choe, Y. J., and Veitch, V. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Qian, C., Liu, D., Wen, H., Bai, Z., Liu, Y., and Shao, J. Demystifying reasoning dynamics with mutual information: Thinking tokens are information peaks in LLM reasoning. In *Advances in Neural Information Processing Systems*, 2025.

Shi, D., Han, Z., Ostermann, S., Jin, R., van Genabith, J., and Xiong, D. Why does reinforcement learning generalize? A feature-level mechanistic study of post-training in large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, 2026.

Sinii, V., Balagansky, N., Gerasimov, G., Laptev, D., Aksenov, Y., Kurochkin, V., Gorbatovski, A., Shaposhnikov, B., and Gavrilov, D. Small vectors, big effects: A mechanistic study of RL-induced reasoning via steering vectors. *arXiv preprint arXiv:2509.06608*, 2025a.

Sinii, V., Gorbatovski, A., Cherepanov, A., Shaposhnikov, B., Balagansky, N., and Gavrilov, D. Steering LLM reasoning through bias-only adaptation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 9202–9211, 2025b.

Team OLMo, Ettinger, A., Bertsch, A., Kuehl, B., Graham, D., Heineman, D., Groeneveld, D., Brahman, F., Timbers, F., Ivison, H., et al. Olmo 3. *arXiv preprint arXiv:2512.13961*, 2025.

Troitskii, D., Pal, K., Wendler, C., McDougall, C. S., and Nanda, N. Internal states before wait modulate reasoning patterns. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 18640–18649, 2025.

Turpin, M., Michael, J., Perez, E., and Bowman, S. R. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, 2023.

Venhoff, C., Arcuschin, I., Torr, P., Conmy, A., and Nanda, N. Understanding reasoning in thinking language models via steering vectors. *arXiv preprint arXiv:2506.18167*, 2025.

Venhoff, C., Arcuschin, I., Torr, P., Conmy, A., and Nanda, N. Base models know how to reason, thinking models learn when. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Wang, C., Feng, Y., Chen, D., Chu, Z., Krishna, R., and Zhou, T. Wait, we don't need to "wait"! Removing thinking tokens improves reasoning efficiency. *arXiv preprint arXiv:2506.08343*, 2025.

Wang, K., Variengien, A., Conmy, A., Shlegeris, B., and Steinhardt, J. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023.

Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., Ren, W., Arulraj, A., He, X., Jiang, Z., et al. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems*, volume 37, pp. 95266–95290, 2024.

Ward, J., Lin, C., Venhoff, C., and Nanda, N. Reasoning-finetuning repurposes latent representations in base models. *arXiv preprint arXiv:2507.12638*, 2025a.

Ward, J., Riechers, P., and Shai, A. Rank-1 LoRAs encode interpretable reasoning signals. *arXiv preprint arXiv:2511.06739*, 2025b.

Yan, G., Sun, C.-E., and Weng, T.-W. ReflCtrl: Controlling LLM reflection via representation engineering. *arXiv preprint arXiv:2512.13979*, 2025.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Yue, Y., Chen, Z., Lu, R., Zhao, A., Wang, Z., Song, S., and Huang, G. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In *Advances in Neural Information Processing Systems*, 2025.

Zhao, R., Meterez, A., Kakade, S., Pehlevan, C., Jelassi, S., and Malach, E. Echo chamber: RL post-training amplifies behaviors learned in pretraining. In *Second Conference on Language Modeling*, 2025.

Zhu, H., Zhang, Z., Huang, H., Su, D., Liu, Z., Zhao, J., Fedorov, I., Pirsiavash, H., Sha, Z., Lee, J., et al. The path not taken: RLVR provably learns off the principals. *arXiv preprint arXiv:2511.08567*, 2025.

## A. Probe Hyperparameter Robustness

**Full $p \times k$ sweep.** Table 3 reports per-layer probe AUC with 95% cluster-bootstrap CIs across pre-break offsets $p-1$ to $p-10$ on Qwen3-8B-Base. Bold marks the discrete pre-emission peak at $L_{22}$ and $L_{30}$. At $L_{22}$ and $L_{30}$ the $p-1$ CI does not overlap the $p-2$ CI (a discrete step); at $L_{10}$ the CIs overlap and the curve is a smooth ramp. Beyond $k=2$ the AUC decays slowly with distance at every layer.

*Table 3.* Pre-emission probe AUC with 95% cluster-bootstrap CIs ($B=1000$) across layer $\times$ pre-break offset $k$ on Qwen3-8B-Base.

| $k$ | $L_{10}$ AUC [95% CI] | $L_{22}$ AUC [95% CI] | $L_{30}$ AUC [95% CI] |
|---|---|---|---|
| 1 | 0.857 [0.849, 0.864] | **0.908 [0.902, 0.914]** | **0.900 [0.893, 0.906]** |
| 2 | 0.851 [0.843, 0.859] | 0.882 [0.875, 0.889] | 0.855 [0.847, 0.863] |
| 3 | 0.849 [0.841, 0.856] | 0.883 [0.876, 0.890] | 0.857 [0.849, 0.864] |
| 4 | 0.848 [0.840, 0.856] | 0.883 [0.876, 0.889] | 0.861 [0.853, 0.867] |
| 5 | 0.842 [0.833, 0.849] | 0.879 [0.872, 0.885] | 0.860 [0.853, 0.868] |
| 6 | 0.836 [0.828, 0.844] | 0.874 [0.867, 0.881] | 0.853 [0.845, 0.860] |
| 7 | 0.827 [0.818, 0.835] | 0.864 [0.856, 0.871] | 0.842 [0.833, 0.850] |
| 8 | 0.822 [0.812, 0.830] | 0.857 [0.849, 0.864] | 0.830 [0.822, 0.838] |
| 9 | 0.812 [0.802, 0.821] | 0.852 [0.844, 0.859] | 0.828 [0.819, 0.836] |
| 10 | 0.814 [0.804, 0.822] | 0.845 [0.837, 0.853] | 0.821 [0.813, 0.830] |

## B. Difficulty Stratification of the Compositional Ablation

We re-ran the headline compositional ablation of §6 stratified by the MATH-500 difficulty level supplied with the dataset. The qualitative pattern is the same at every level: $\mathbf{v}_{\mathrm{BT}}$ ablation lowers pivot rate, $\mathbf{w}_E$ ablation slightly raises it, and the joint condition tracks the linear-additive prediction within seed noise. The sign of every cell of Table 2 is preserved across levels; the magnitude of the $\mathbf{v}_{\mathrm{BT}}$ drop is somewhat larger on harder problems, consistent with harder problems eliciting more pivots in the baseline. Final-answer accuracy is preserved at every level. We do not report the per-level numbers in the body because the per-stratum sample size is small and the across-stratum variation is within the cluster-bootstrap CI of the aggregate.

## C. Out-of-Corpus Probe Transfer (AIME 2022–2025)

The headline $L_{22}$ pre-emission probe is fit on MATH-500 and Venhoff-500 traces, both of which Qwen3-8B may have seen during pretraining. To rule out that the AUC of $0.908$ is an artifact of contamination, we apply the probe *without retraining* to Qwen3-8B-RL traces of AIME 2022, 2023, 2024, and 2025 problems. AIME 2025 in particular post-dates the pretraining cutoff and is the cleanest single-year transfer test.

The probe transfers cleanly. Combined across the four contest years, the $L_{22}$ AUC is $0.903$, within $0.5$ pp of the in-corpus headline. The 2025-only subset alone gives $L_{22}$ AUC $0.847$. Because each AIME year is a small set, we report the combined number as the primary defense and the 2025-only number as the post-cutoff lower bound. The pre-emission cliff is also preserved: the AUC at $p-1$ is separated from the AUC at $p-2$ by a step whose CIs do not overlap, just as in the in-corpus probe.

## D. Qwen3-32B Scale Replication

At Qwen3-32B (64 layers, $\sim 4\times$ the parameter count of the 8B), the per-layer pre-emission probe peaks in the mid-to-late stack at the same relative depth as on Qwen3-8B; Figure 2b shows the curve. Sensor projection-out at the peak layer drops pivot rate by approximately $10$ pp across two seeds, with accuracy preserved and a random-vector control flat. The sensor side of the mechanism therefore scales.

The evaluator side is more fragile at this scale. The clean cosine-orthogonal evaluator location (§6 discussion) sits at $L_{62}$ rather than at the linear-probe peak, which is contaminated by the sensor. Even at the orthogonal location, the size and sign of the $\mathbf{w}_E$ effect are not robust across the two seeds at the pilot $N=25$ used here, with the seed-mean compositional super-additivity passing through zero. We therefore report the Qwen3-32B circuit-side findings as a corroboration of the sensor universality and as evidence that the clean evaluator location must be located within-layer at scale, not as a full circuit replication.

## E. OLMo-3-7B-Think Cross-Family Replication

OLMo-3-7B-Think is a Llama-derived 32-layer reasoning model in a different post-training family from Qwen3. We re-derived an OLMo-native pivot lexicon by inspecting the empirical distribution of first-words after paragraph breaks in OLMo's own generations, since OLMo uses a partially different rhetorical-pivot vocabulary. With native labels, the pre-emission probe peaks in the mid-to-late stack at the same relative depth as on Qwen3 and reaches AUC roughly $0.82$ at the peak layer. Sensor projection-out during generation lowers pivot rate by a few percentage points with accuracy preserved.

The full four-part circuit is not a clean replication on OLMo. The evaluator probe at the analogous mid-late location separates correct from incorrect trajectories at moderate AUC, but the sign and magnitude of the compositional gate effect are smaller and noisier than at Qwen3-8B and do not yield a stable opposite-signed reading at the pilot $N{=}50$–$100$ used here. We therefore read the OLMo evidence as confirming the universal pre-emission sensor and the decorative-for-accuracy claim cross-family, but not as a re-derivation of the gate.

## F. Comparison with Related Backtracking-Direction Work

The closest precedent to our pre-emission analysis is Troitskii et al. (2025), who probe the residual stream immediately before each `wait` token via sparse crosscoders and identify features that promote and suppress wait emission. We share the pre-emission framing and extend it in three ways: a single linear direction with a per-token cliff at $p{-}1$ (their analysis is at the feature-graph level), a causal isolation under generation-time projection-out, and a four-part decomposition of the surrounding circuit including the RL-installed trigger.

Venhoff et al. (2026) and Venhoff et al. (2025) extract base-model directions for backtracking, verification, hypothesis-checking, knowledge-retrieval and uncertainty-management, and recover up to $91\%$ of the base-to-thinking accuracy gap by activating the five directions in concert. Our $\mathbf{v}_{\mathrm{BT}}$ direction is the backtracking arm of that composite. The novel contributions on top of that line are mechanism-level isolation of one arm, evidence that the isolated arm is decorative for accuracy, and the four-part circuit decomposition that splits the gating mechanism away from the underlying directions.

Ward et al. (2025a) and Yan et al. (2025) steer reflection / reasoning behaviors via representation engineering on the same family of post-trained models. Our protocol differs in using projection-out (not addition) so that we measure the causal effect of removing a signal rather than the response to perturbing it, and in evaluating accuracy under the intervention rather than the rate of the targeted behavior alone. Wang et al. (2025) is the closest behavioral analogue: they suppress `Wait`/`Hmm` tokens at the logit level and report unchanged accuracy across five model families. Our isolation result extends that effect to the residual stream: the upstream signal that drives those tokens is itself causally inert for accuracy.

## G. Pivot/Continuation/Enumeration Labelling

A paragraph break is any token position $p$ in a generated trace matching the regular expression `[.?!]\n(\n?)` immediately following a content token. Each break is labelled by the first content token of the next paragraph using a deterministic three-way classifier.

**Pivot openers.** On Qwen3 and Qwen3-32B, the pivot lexicon contains `Wait`, `Alternatively`, `Hmm`, `Actually`, `But`, `However`, `Hold on`, and the lower-cased variants where they occur as paragraph openers. The list was assembled by inspecting the empirical distribution of first-words after breaks in held-out Qwen3-RL traces and freezing the head of that distribution that is consistent with rhetorical hesitation.

**Enumeration openers.** Numeric or bullet markers at the start of the next paragraph: `1.`, `2)`, `(a)`, `Step 1`, `First`, `Second`, the dash and asterisk bullets, and similar regex-detectable list openers. Enumeration is treated as a separate structural mode from pivot in the three-way probe of §6.

**Continuation.** Any first-token that is not a pivot or enumeration opener.

**OLMo-native lexicon.** OLMo-3-7B-Think uses a partially different rhetorical-pivot vocabulary; its pivot lexicon was re-derived by the same procedure on held-out OLMo generations. Without that re-derivation, the standard Qwen3 lexicon under-counts pivots on OLMo.

## H. Hardware and Runtime

All experiments run on a single workstation with $4\times$ NVIDIA A100 80 GB GPUs (CUDA 12.x), $256$ GB of host RAM, and an AMD EPYC CPU. Qwen3-8B-Base/RL and OLMo-3-7B-Think experiments use 1–2 GPUs in fp16; Qwen3-32B uses 4 GPUs with tensor parallelism. The pre-emission probe (§3) is fit on CPU. Total wall-clock for the headline pipeline (probe extraction, AIME defense, all $4$-seed compositional ablations on Qwen3-8B, layerwise activation-patch sweep, six trigger-localization rungs) is approximately $180$ GPU-hours; the Qwen3-32B and OLMo-3-7B-Think replications add roughly $80$ and $40$ GPU-hours respectively. End-to-end disk usage for cached residuals and traces is $\sim 200$ GB. Model weights are pulled from public Hugging Face checkpoints at runtime.