Backtracking is Decorative

Removing the signal that drives "Wait, …" halves backtracking — and leaves accuracy unchanged

Unai Zulaika · Miguel Fernandez-de-Retana · Rubén Sánchez-Corcuera · Aritz Bilbao-Jayo · Aitor Almeida UNIV. OF DEUSTO, BILBAO

Deusto
Universidad de Deusto
Deustuko Unibertsitatea
University of Deusto

CODE

WHAT WE DID

Isolated the residual-stream signal behind "Wait, …" pivots in **three reasoning families** — Qwen3-8B, Qwen3-32B, OLMo-3-7B-Think — projected it out **during generation**, then mapped the full circuit on Qwen3-8B with probes, RL→Base activation patching, and a six-rung weight-localization ladder.

WHAT WE FOUND

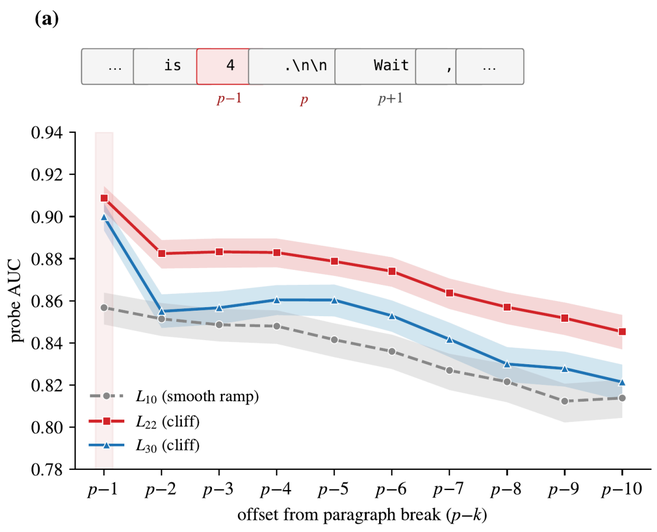
The pivot decision is settled **one token before the break** by a pretraining-native direction. Removing it cuts backtracking **21–78%** with accuracy flat — math and non-math. Of the four-part circuit, **three parts precede RL**; RL installs only the **trigger**, and even that is diffuse.

WHY IT MATTERS

The most visible RL-installed reasoning behavior is not the channel through which RL improves accuracy. Decomposing it into mostly pretraining-native parts points toward eliciting reasoning by targeted manipulation of existing structure, rather than full RL post-training.

01 The decision precedes the token

PRE-EMISSION PROBE



A discrete cliff, not a distance ramp. At L22/L30 the AUC at $p-1$ is a step the $p-2$ CI cannot reach (−2.6 / −4.5 pp); at L10 the curve is a smooth proximity ramp. The decision becomes a discrete state at the last pre-break token, mid-late stack.

A logistic probe at `(L22, p-1)` reads the upcoming pivot at **AUC .908** [.902–.914] — leave-one-trace-out over 1,445 traces / 13,299 breaks — and transfers Base↔RL within noise: **the sensor is pretraining-native.**

SAME SENSOR, THREE FAMILIES	PEAK	REL. DEPTH	AUC
Qwen3-8B-RL	L23/35	0.66	.91
OLMo-3-7B-Think	L17/31	0.55	.82
Qwen3-32B	L47/63	0.75	.88

02 Pivots are decorative for accuracy

CORE FINDING · 3 CONVERGING LINES

DEFINITION

A behavior is **decorative for accuracy** if removing the residual-stream signal that drives it produces a large change in the behavior's rate together with **no detectable change in accuracy** under the same generation protocol.

CAUSAL

−59%

pivot rate under sensor projection-out on Qwen3-8B (28.0→11.4%); accuracy 80→83%; random-vector control flat at baseline

OBSERVATIONAL

r = +.03

per-trace pivot rate vs. correctness, MATH-500 (N=225); non-math r = −.15 — both 95% CIs cross zero

CROSS-MODEL

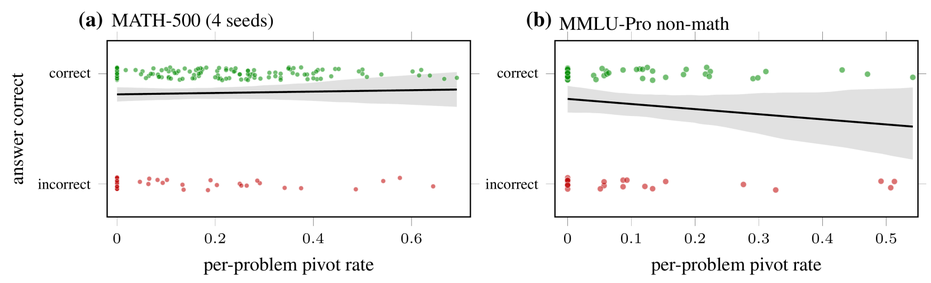
+22 pp

OLMo-3-7B-Think pivots ~1.5× *less* than Qwen3-8B-RL on MATH-500, yet scores ~22 pp *higher*

SENSOR PROJECTION-OUT	LAYER	N×SEEDS	PIVOT RATE	Δ	ACCURACY
Qwen3-8B-RL	L22/35	100×4	28.0→11.4%	**-16.6pp**	80→83%
OLMo-3-7B-Think	L17/31	50×3	20.5→16.1%	**-4.4pp**	82→84%
Qwen3-32B	L47/63	25×2	13.4→3.0%	**-10.4pp**	80→74%

Accuracy within a few pp of baseline in every condition × seed cell; the 32B run is pilot-N (25×2). Same problem, with and without the circuit:

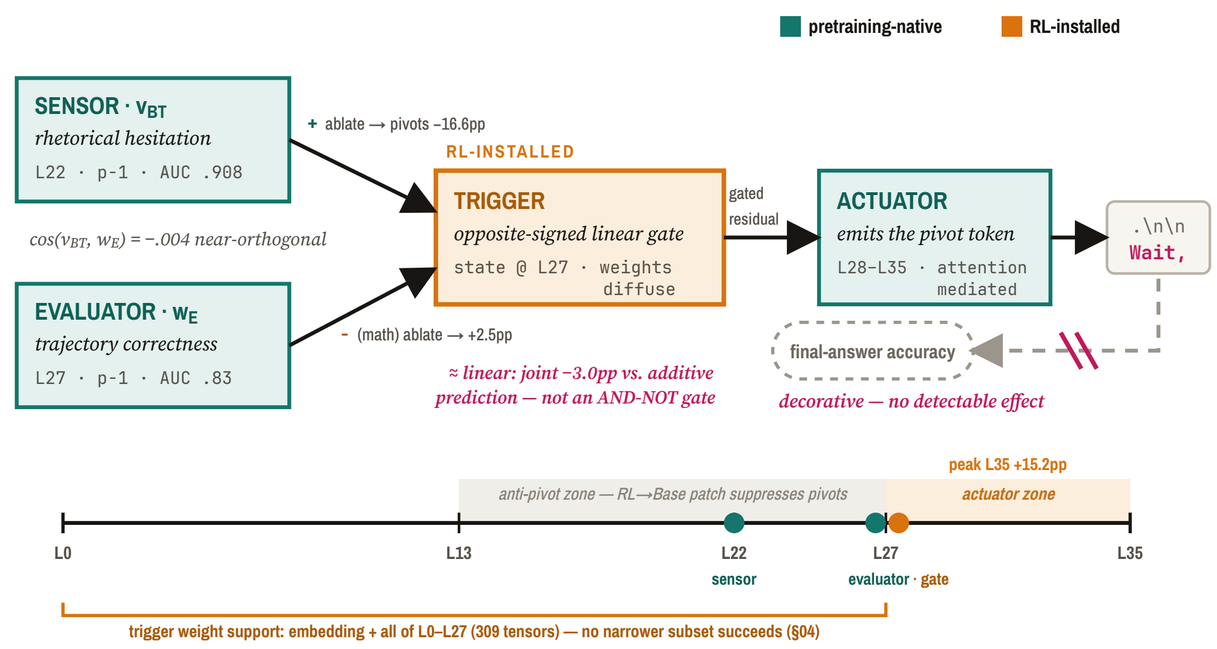
```
BASELINE  …so x=4. \n\n Wait, let me verify: 4·3=12  →  final: x=4
ABLATED   …so x=4. \n\n Therefore,  →  final: x=4  — no pivot, same answer
```



Per-problem pivot rate is statistically independent of getting the right answer — in un-intervened generation as in ablation, math (left) and non-math (right). **Scope:** this rules out backtracking-*by-itself* as RL's accuracy channel, not backtracking-in-composite with the other steering arms of Venhoff'25.

03 One map: the pivot circuit at Qwen3-8B

SENSOR · EVALUATOR · TRIGGER · ACTUATOR



COMPONENT	IDENTIFIED BY	CAUSAL EVIDENCE
Sensor `$v_{BT}$ · L22, p-1`	pivot–continuation centroid; probe AUC .908, cliff at p−1	project out → pivots −16.6pp, accuracy flat; random-vector control flat
Evaluator `$w_E$ · L27, p-1`	correct–incorrect centroid; AUC .83 in Base & RL (Δ +.009)	project out → pivots +2.5pp on math; sign flips by domain (§05)
Trigger `@ L27 · diffuse`	full-L27 RL→Base activation patch recovers RL's pivot rate (27% vs. 11% floor)	six narrow weight transplants null; embed + L0–L27 → 60% of ceiling (§04)
Actuator `L28–L35`	per-layer RL→Base patch sweep	promotes pivots toward RL ceiling, peak L35 +15.2pp; works in Base

COMPOSITIONAL ABLATION (100×4)	PIVOT RATE	Δ	ACCURACY
baseline	28.0%	—	80%
ablate v_{BT} at L22	11.4%	-16.6pp	83%
ablate w_E at L27	30.5%	+2.5pp	79%
ablate both	11.0%	-17.0pp	81%

The two pretraining-native directions combine **approximately linearly with opposite signs** — joint ablation lands −3.0pp *below* the additive prediction, consistently across seeds. An N=25 pilot had said +12.9pp super-additive (an AND-NOT story); it did not survive N=100 × 4 seeds.

Three of four components precede RL (teal); RL contributes only the trigger (orange). Layer ruler from the per-layer RL→Base patch sweep: copying RL's $p-1$ residual into Base at L13–L27 *suppresses* pivots below Base's floor; at L28–L35 it *promotes* them toward RL's ceiling — the actuator is in Base already, and fires whenever the L27 state is set. **v_{BT} tracks rhetorical hesitation, not failure:** logit-lens top lexemes `anyways · anyway · Maybe`; orthogonal to content and enumeration.

MORElab · DeustoTech

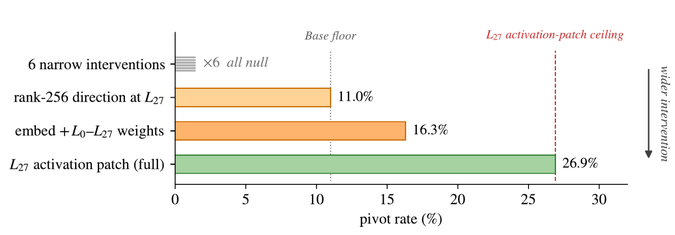
04 RL's contribution is real — and diffuse

LOCALIZATION LADDER

Patching the full L27 residual from RL into Base recovers RL's native pivot rate (26.9% vs. Base's ~11% floor). Holding that ceiling fixed, we walk progressively wider interventions:

SIX NARROW INTERVENTIONS — ALL NULL AT THE BASE FLOOR

rank-K direction injection at L27, K ≤ 128 · single head `L35.H12` · L27 attention-only / MLP-only / both sublayer patches · single-layer L27 weight transplant · L22–L27 band transplant



The trigger has no narrow weight support. Only wide scope recovers pivot mass: rank-256 reaches the Base floor; the smallest succeeding *weight* subset is the **embedding plus all of L0–L27** (309 tensors), at ~60% of the ceiling. RL made small adjustments across the whole upstream pipeline; the L27 state they produce is what the pretraining-native actuator reads — the sparse-but-distributed RLVR picture (Meng'26; Zhu'25).

v_{BT} and w_E are clean linear-representation objects; **the trigger is not** — high-rank in activations, spread in weights. Single-direction analyses can identify the gate's *inputs* but not the gate. Our nulls are at the granularities tested: *failing to localize is not proof of absence.*

05 Predicting the gate's sign across domains

PRE-REGISTERED · 6/6

The circuit is pinned on math. Does it generalize? We make it falsifiable: the *sign* on which w_E (evaluator) enters the gate should track a domain's error structure, not the math distribution.

Natural-trace asymmetry **(observed)** predicts the ablation sign **(causal)** — no intervention:

DOMAIN	TRAJECTORY STATE · P−1	w_E → GATE	ABLATE w_E
RESCUE · math pivoting flags trouble	*incorrect* states pivot more	opposite-signed	**pivots ↑** +2.5pp
REVISION · law · psych · econ pivoting is careful revision	*correct* states pivot more	same-signed	**pivots ↓**

PRE-REGISTERED

6/6

sign predictions correct, every MMLU-Pro domain tested

Signs predicted from the **natural-trace correctness asymmetry alone** — economics was called **before** its ablation was run, and every sign held when replicated at higher N.

The recipe: read a domain's error–pivot asymmetry, predict the gate sign, and extend the mechanism cross-domain *without re-running the causal protocol.*

WHERE THE ACCURACY ACTUALLY LIVES

If backtracking is decorative, the accuracy gap lives elsewhere. Composite steering of five base directions recovers up to **91%** of the base-to-thinking gap (Venhoff'25); with backtracking inert in isolation, that capacity sits in the other four arms — **verification, hypothesis-checking, knowledge-retrieval, uncertainty-management**. Mapping each is the immediate program.

TAKEAWAYS FOR PRACTITIONERS

- → **Surface-token suppression is now mechanism-level.** Projecting one pretraining-native direction out halves visible backtracking at zero accuracy cost — extends the logit-level result of Wang'25 to the residual stream, and gives a token-budget knob.
- → **Opposing-sign inhibitors mask steering.** An injection at L27 looked dead — only because w_E was canceling it; remove w_E and it drives pivots. A null result can be a false negative: re-test with the inhibitor out.
- → **If pivots don't carry accuracy, what do they carry?** Possibly legibility — backtracking as interface for the reader, not computation for the model.

LIMITATIONS

The full four-part circuit is confirmed in only one model (Qwen3-8B-RL) on math; in OLMo and 32B the evaluator and trigger rest on small, noisy samples (the sensor and the decorative result do replicate). 'Decorative' means backtracking *alone* doesn't change accuracy — we haven't ruled out that it helps when combined with the other reasoning behaviors. In 32B the usual evaluator probe gets mixed with the sensor, so it had to be isolated a different way (a within-layer search at L62). The trigger's null results only cover the scales we tested; a finer scale might still find it.

Venhoff+ '25 · Wang+ '25 · Troitskii+ '25 · Ward+ '25 · Meng+ '26 · Zhu+ '25 · Wang+ '22 (IOI) · McDougall+ '23 · Park+ '23