Ghost Heads Across Training: When Greedy and Distributional Patching Disagree

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

Abstract

A language model can produce a correct answer within 32 samples long before it does so reliably on the first try. Across arithmetic, GSM8K, and MATH500, pass@32 saturates during pretraining while pass@1 lags far behind. What mechanistic process closes this gap? We track circuit formation across OLMo-3-7B's full training pipeline (pretraining, mid-training, and RL-Zero) with both greedy and distributional activation patching, and introduce *answer-token patching* for multi-step reasoning. The two metrics agree on most components but *dissociate* at specific attention heads, which we call *ghost heads*: heads with high greedy recovery but near-zero distributional impact. Ghost heads are a recurring training instability: they peak whenever the training objective shifts and decay within each phase. Mid-training resolves the dissociation: patching effects concentrate onto a few attention heads where both metrics agree. RL-Zero, which improves pass@1 without touching pass@32, does not reintroduce ghost heads; two independent tests (noise sensitivity and verifier reranking) confirm it acts as a diffuse whole-model shift rather than reorganizing individual circuits. The phenomenon replicates across three model families.

1. Introduction

A 7B language model can solve 98% of 3-digit addition problems within 32 samples at a training stage where its single-sample accuracy is 38%. What changes between these two modes of performance? Recent work argues that post-training, including RL-Zero, primarily reshapes how existing capability is sampled rather than expanding it (Yue et al., 2025; Zhao et al., 2025; Wu et al., 2025; Zhu et al., 2025). We confirm this behavioral pattern across 14 OLMo-3-7B checkpoints (§3.1) and take it as our starting point: what is the mechanistic substrate of this *reliability gap*, and what does each training stage change about it?

Mechanistic interpretability offers several such tools: activation patching localizes causal contributions of individual components (Meng et al., 2022; Zhang & Nanda, 2024); weight transplantation tests functional decomposability; weight-diff analysis characterizes the geometry of parameter changes. Several prior works track circuits across training (Tigges et al., 2024; Nikankin et al., 2025; Hakimi et al., 2025), but two gaps remain. First, standard activation patching degrades on multi-step reasoning, where single-operand corruption barely moves the first-token prediction. Second, prior temporal work reports a single patching metric (greedy logit recovery), leaving open whether the observed circuit dynamics are robust to distributional metrics that measure probability-mass changes rather than argmax changes.

We address both gaps. We trace circuit formation across 14 OLMo-3-7B checkpoints covering pretraining, mid-training, and RL-Zero, using activation patching with both greedy and distributional metrics, weight transplantation, weight-diff geometry, and a new method called *answer-token patching* for multi-step reasoning. Our central finding extends Zhang & Nanda (2024)'s metric-dissociation analysis temporally: the disagreement between greedy and distributional patching is not a static property of attention heads but a non-stationary signature that peaks at training-objective shifts and resolves into convergence at the base model. We call the heads exhibiting this dissociation *ghost heads*, identify L31.H13 as the canonical example, and show its cross-model analogue is the only head out of 1024 whose dissociation score exceeds chance under a sign-flip permutation test (§4). On the RL-Zero transition, two independent MI probes and five weight-dependent measurements return null or structurally symmetric results, consistent with an argmax/sampling-layer account of RL-Zero proposed by Yue et al. (2025); Zhao et al. (2025); Wu et al. (2025); Zhu et al. (2025), in which RL-Zero acts as a whole-model distributional shift without an identifiable head-level locus.

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

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

Contributions.[1]

1. *Ghost heads* (§3.2): we identify attention heads whose greedy recovery is not matched by distributional impact. The dissociation peaks at training-objective shifts and replicates across three architectures.

2. *Answer-token patching* (§5): a method that extends causal circuit analysis to multi-step reasoning. We find no evidence that RL improves verification ability on any of three tasks.

3. *RL-Zero null* (§4): two independent probes and five weight-dependent measurements find no head-level signature of the RL-Zero transition on in-domain math. Cross-model patching links the canonical ghost head L31.H13 to the base-to-RL behavioral delta ($p<0.001$, sign-flip permutation test over 1024 heads).

2. Experimental Setup

Models. We study OLMo-3-7B (OLMo Team, 2025) across 14 checkpoints: four from stage-1 pretraining (5.93T tokens), three from stage-2 mid-training (100B curated tokens, 19% math), three from stage-3 long-context training, the final base model, and three from RL-Zero-Math post-training (steps 200, 1000, 1900). For cross-model validation we use OLMo-2-7B (OLMo Team et al., 2025) (4 checkpoints) and, for cross-architecture validation, DeepSeek-Math-7B and Qwen2.5-Math-7B at their base checkpoints.

Tasks. Our primary task is 3-digit addition with carries: we sample 200 operand pairs (a, b) uniformly from $[100, 999]^2$, rejecting cases without any carry (seed $= 42$). Each clean prompt has the form "`Q: What is` a + b`?\nA:`"; the corrupted variant replaces b with a different random 3-digit integer b', giving the incorrect target $a + b'$. The clean/corrupt pair enables symmetric activation patching. We extend to GSM8K (1296 patchable word problems) and MATH500 (679 patchable competition-math problems).

Activation patching with dual metrics. We follow Meng et al. (2022) with symmetric token replacement (Zhang & Nanda, 2024; Wang et al., 2023); we treat patching as suggestive rather than fully scrubbed (Geiger et al., 2025; Chan et al., 2022). For each component c (MLP layer, attention layer, or individual head) we compute two recovery metrics:

$$g(c) = \frac{\ell_c^{\text{patch}} - \ell^{\text{corrupt}}}{\ell^{\text{clean}} - \ell^{\text{corrupt}}}, \qquad p(c) = \frac{\log \pi_c^{\text{patch}} - \log \pi^{\text{corrupt}}}{\log \pi^{\text{clean}} - \log \pi^{\text{corrupt}}} \tag{1}$$

where ℓ is the correct-answer logit and π its softmax probability. Both equal 0 at no recovery and 1 at full recovery; the ghost-head signature is $g(c) \gg p(c)$. We use trimmed means (10% each tail) across examples for distributional metrics (Zhang & Nanda, 2024; Heimersheim & Nanda, 2024) and the Herfindahl-Hirschman Index (HHI) to quantify circuit concentration.

Answer-token patching. Standard activation patching localizes circuits on single-token-output tasks, but it degrades on multi-step reasoning: corrupting one operand of a chain-of-thought problem barely shifts the first-token prediction, leaving fewer than half of GSM8K and MATH500 examples with a usable clean–corrupt gap (Appendix D). This restricts standard temporal patching analyses, including the ghost-head trajectory we report below, to single-token tasks such as 3-digit arithmetic. We introduce *answer-token patching* as a workaround: present the model with a problem and its complete chain-of-thought solution, corrupt only the final numerical answer, and patch on the token predicting "correct" or "incorrect" (see Appendix K for worked example prompts). This converts multi-step reasoning verification into a single patchable forward pass that retains a strong clean–corrupt signal. We validate the method on arithmetic, where it recovers the same specialist head L29.H5 found by standard patching, then apply it to GSM8K and MATH500 (§5).

Weight analysis. We construct hybrid models by swapping MLP or attention weights between checkpoints (weight transplantation), analyze weight differences ΔW via Frobenius norms and SVD decomposition across training stages, and test perturbation sensitivity via isotropic Gaussian noise at varying scales (Appendix I).

3. Ghost Heads Across Training

3.1. The Reliability Gap

We evaluate OLMo-3-7B across all 14 checkpoints on pass@k ($k \in \{1, 32\}$, T=0.8, top-p=0.95) for arithmetic, GSM8K, and MATH500 (Figure 1, Table 2). Pass@32 saturates during pretraining on all three tasks while pass@1 lags by 25–50 percentage points. Mid-training closes the gap; RL-Zero peaks pass@1 briefly (step 200) then degrades it while pass@32 is preserved on all three tasks, ruling out capability destruction. Each task recruits a different specialist attention head (L18.H7 for arithmetic, L26.H31 for GSM8K,

[1] Code and data are available at https://github.com/unai-zulaika/ghost-heads.

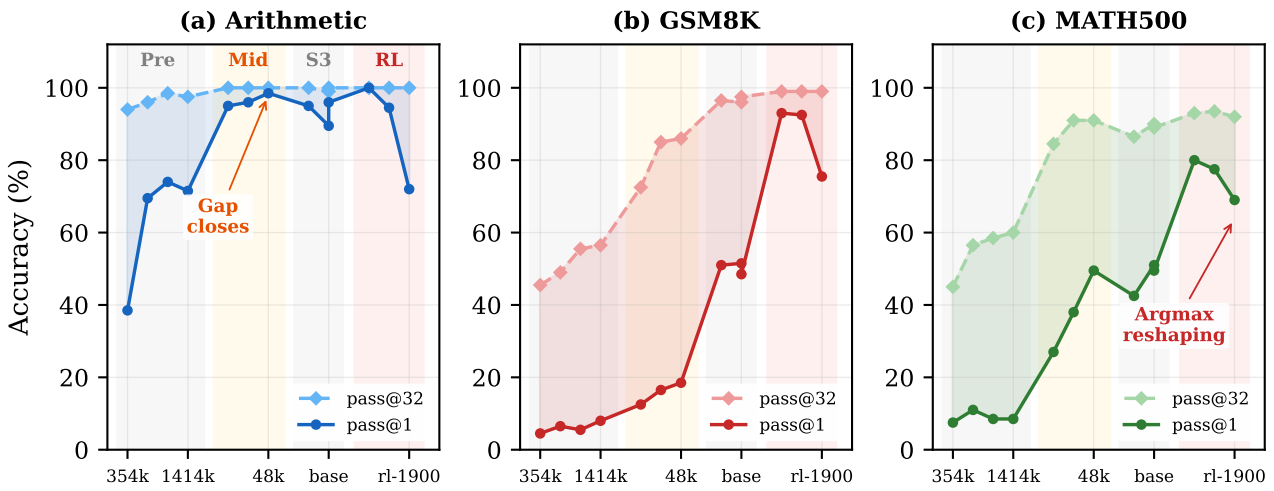

Figure 1. **The reliability gap across tasks.** Pass@1 (solid) and pass@32 (dashed) on arithmetic (a), GSM8K (b), and MATH500 (c) across 14 OLMo-3-7B checkpoints. Pass@32 saturates during pretraining while pass@1 lags. Mid-training closes the gap. Late RL degrades pass@1 while pass@32 is preserved on all three tasks, consistent with argmax reshaping rather than capability loss.

Table 1. **Cross-task summary.** Each task recruits a different specialist head while sharing MLP-dominant computation. HHI is computed both on the full set (*all*) and on the high-signal subset where the clean-corrupt logit gap is positive (*filtered*); the gradient from arithmetic to MATH500 strengthens under filtering, indicating that the original gradient was not a denominator artifact at low signal quality. Signal quality is the fraction of examples passing the filter. The top MLP layer is mlp_31 for single-step arithmetic, GSM8K, and MATH500; mlp_22 dominates multi-step arithmetic computation (Appendix C).

	Arithmetic	GSM8K	MATH500
Specialist head	L18.H7	L26.H31	L20.H21
HHI (greedy, all)	0.111	0.026	0.015
HHI (greedy, filtered)	**0.152**	**0.045**	0.013
Top MLP (single-step)	mlp_31	mlp_31	mlp_31
Signal quality	51%	36%	44%
n (filtered)	166	294	255

L20.H21 for MATH500) while sharing MLP-dominant computation. MLP layer 31 is the top contributor across all three tasks (Table 1).

3.2. Ghost Heads: Definition and Mechanism

We identify ghost heads on the 3-digit arithmetic task using standard activation patching with both greedy and distributional metrics; we then verify in §5 that the same dominant head L18.H7 and the same secondary head L29.H5 are recovered by answer-token patching on multi-step problems, providing an independent confirmation of their causal roles. MLP layers 30–31 dominate patching recovery across all checkpoints, providing a stable computational substrate. Their internal neurons rewire during pretraining ($\sim$20% overlap) then freeze through RL ($>$96% overlap). The attention circuit shows different dynamics that depend critically on the patching metric.

Table 2. **Behavioral trajectory.** Pass@1 / pass@32 (%) at four representative checkpoints. RL transiently peaks pass@1 then degrades it while pass@32 is preserved across all three tasks.

	pass@1 / pass@32			
Task	end-S1	base	RL peak	RL end
Arithmetic	72 / 98	96 / 100	100 / 100	72 / 100
GSM8K	8 / 57	49 / 98	93 / 99	76 / 99
MATH500	9 / 60	50 / 89	80 / 93	69 / 92

Hypotheses. We test four predictions: **H1:** distributional patching should yield systematically lower recovery than greedy for specific heads (dissociation, not uniform scaling); **H2:** ghost-head count should peak at training-regime boundaries and decay within each phase (training instability, not fixed architecture); **H3:** MLPs should not exhibit ghost signatures: softmax forces every attention head to contribute regardless of alignment with the correct-token direction, decoupling routing from contribution; MLP neurons have no analogous normalization constraint; **H4:** the top ghost head's greedy-to-distributional recovery ratio (g/p) should be robust to bootstrap resampling and exceed chance under a sign-flip permutation test, distinguishing causal misdirection from measurement noise.

Under greedy patching, three attention heads alternate as the top contributor by greedy recovery $g(c)$ across training stages: L31.H13 dominates in early pretraining, L29.H5 at the end of pretraining, and L18.H7 at the base checkpoint (Figure 2a). Distributional patching dissolves this relay:

L18.H7 dominates throughout, L31.H13 has near-zero probability recovery for its entire trajectory, and the apparent circuit transfer is an artifact (Figure 2b), confirming H1.

We call L31.H13 a *ghost head*: a component whose high greedy recovery is not matched by distributional impact. Ghost heads perturb which token wins the argmax without moving probability mass toward the correct answer.

Value-state norms ($\|V \cdot x\|_2$ at the final token) reveal that ghost heads are not suppressed; they are *strong but misdirected*. At stage1-354k, L31.H13 has the largest value norm of the three key heads (13.96 vs. H7's 2.80); ghost heads have $5.93\times$ higher norms than the model average. Their outputs shift which token wins the argmax without aligning with the correct token's probability mass: restoring L31.H13's clean output shifts the argmax 17.6% of the way back (greedy recovery) but moves probability mass only 0.9% (prob recovery), a 19:1 causal dissociation. Training then collapses L31.H13's norm by $19\times$ (to 0.74); at base it meets standard dormancy criteria (Guo et al., 2024; Sandoval-Segura et al., 2026), but our criterion fires at stage1-354k when the head is maximally active and maximally misdirected, precisely the phase dormancy criteria miss (Spearman(value-norm, GhostScore) positive at every checkpoint; Appendix Q).

This admits a deflationary reading: ghost heads may simply perturb logits near the argmax decision boundary, and greedy recovery is sensitive to such boundary perturbations while probability recovery is not. We do not contest this at the single-checkpoint level. What it does not explain is the *temporal structure*: ghost heads peak at objective shifts, decay within each phase, and L31.H13 recurs at every boundary (GhostScore 10.5 at stage1-354k, 9.0 at stage2-31k), supporting H2. The dissociation is architecturally specific to attention (H3): softmax forces every head to contribute regardless of task-alignment; MLP neurons have no such constraint, and we find no MLP ghost signatures at any checkpoint.

Continuous GhostScore. To rank heads without a binary cutoff, we define

$$\mathrm{GhostScore}(h) = \frac{\max(0, g(h))}{\max(|p(h)| + \epsilon,\, \phi)} \tag{2}$$

where ϵ=0.005 stabilizes and ϕ=0.01 floors the denominator. A score $\gg 1$ flags a greedy-dominant dissociation; the qualitative trajectory matches the binary criterion at every threshold (Appendix N). H4 is confirmed in §4 by the cross-model permutation null (p<0.001 over 1024 heads).

3.3. The 14-Checkpoint Trajectory

Applying the ghost-head analysis to all 14 checkpoints reveals a richer picture than the two-checkpoint comparison suggests (Figure 3, Table 3).

Table 3. **Ghost-head trajectory across 14 checkpoints.** Ghost-head count, r(greedy, dist), and per-head recoveries for L31.H13 (canonical ghost head) and L18.H7 (true specialist). g and p are greedy and probability recovery (Eq. 1). Ghost heads peak twice: in early pretraining and mid-stage mid-training.

			L31.H13		L18.H7	
Checkpoint	ghosts	r	g	p	g	p
Stage 1: pretraining						
step354k	**2**	.42	.176	.009	.081	.158
step707k	**4**	.51	.008	.001	.031	.130
step1060k	0	.60	.011	.000	.015	.139
step1414k	0	.80	.038	−.001	.009	.006
Stage 2: mid-training						
step16k	0	**.99**	−.003	.000	**.682**	**.725**
step31k	**4**	**.26**	**.105**	.007	.092	.027
step48k	1	.78	.131	.015	.138	.146
Stage 3: long-context						
step1k	0	.998	.000	.000	.746	.780
step6k	1	.81	.002	.000	−.017	.068
step11k	0	.99	.000	.000	.483	.537
base	0	.998	.000	.000	.478	.488
RL-Zero						
step200	0	.996	.000	.000	.459	.434
step1000	0	.992	.000	.000	.501	.673
step1900	0	.997	.000	.000	.509	.575

Three findings emerge.

Ghost heads are a recurrent instability. Ghost heads peak twice: in early pretraining (stage1-707k: 4 heads) and in mid-stage mid-training (stage2-31k: 4 heads). The two peaks coincide with moments when the training objective changes: the start of pretraining, and the shift from pretraining data to curated mid-training data. Between and after these peaks, ghost heads decay and eventually disappear. The phenomenon is not a monotonic pretraining artifact: it is a recurrent instability tied to training-objective shifts, distinct from the monotone emergence of induction heads (Olsson et al., 2022).

Mid-training is oscillatory, not smooth. The greedy/distributional correlation crashes to $r = 0.26$ at stage2-step31k (the worst point in the entire 14-checkpoint trajectory) and recovers to $r = 0.78$ by step48k (Table 3). The circuit forms, disrupts, and reforms within 48k steps of mid-training, with full stabilization occurring only in stage 3. Even the true specialist L18.H7 is not immune: at stage3-step6k it briefly shows negative greedy (-0.017) but positive distributional recovery (0.068), an inverse-ghost signature that resolves by step11k. Monitoring the greedy/distributional correlation during mid-training could serve as an early diagnostic for circuit

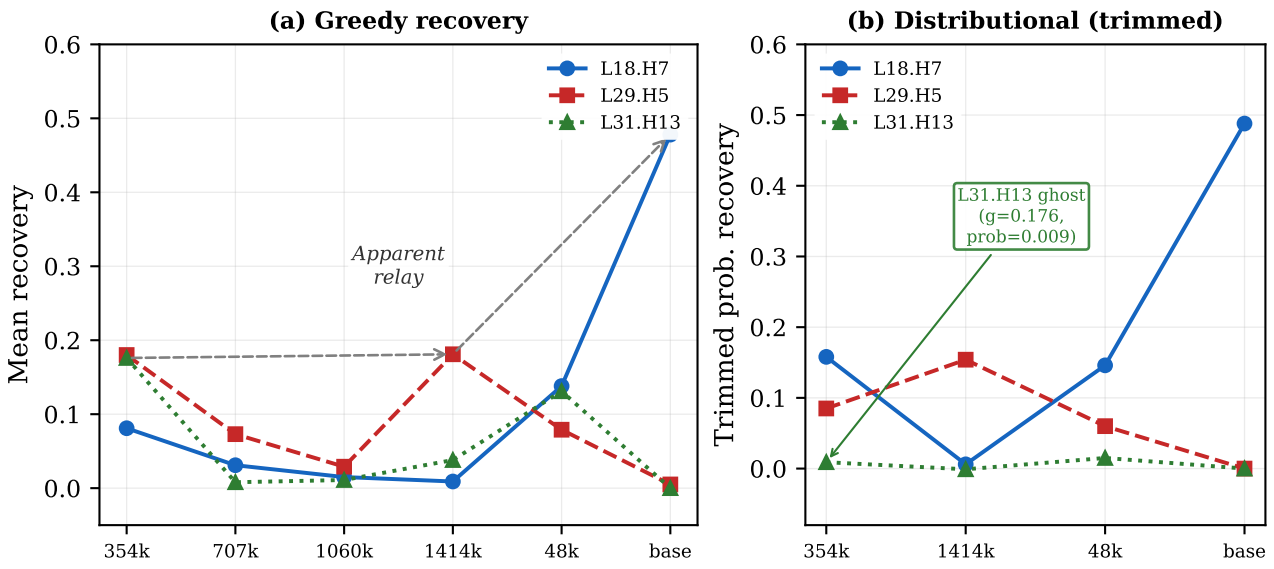

Figure 2. **Ghost heads: greedy and distributional patching diverge.** (a) Greedy recovery shows an apparent relay across three heads. (b) Distributional recovery reveals L31.H13 as a ghost head (greedy 0.18, prob 0.009) and L18.H7 as the true dominant head throughout.

instabilities before they manifest behaviorally.

Mid-training resolves the instability. By the base checkpoint, ghost heads are eliminated: r reaches 0.998, and the specialist head L18.H7 is identified by both metrics with matched greedy (0.478) and distributional (0.488) recovery (Table 3). The resolution occurs primarily during mid-training, where L18.H7 first emerges as the dominant head with both greedy and distributional recovery above 0.68 (step16k). The ghost-head analysis above uses standard single-token patching on arithmetic; §5 extends this to multi-step reasoning via answer-token patching and confirms the same specialist heads on GSM8K and MATH500.

RL-Zero creates zero ghost heads on arithmetic. Across all three RL checkpoints (steps 200, 1000, 1900), the canonical criterion finds zero ghost heads on the 3-digit arithmetic task and r stays above 0.99. RL-Zero's perturbation on this task is too small and diffuse to create the concentrated greedy-without-distributional signal that characterizes ghost heads.

MATH500 non-result. We probed whether L29.H5 develops a ghost signature on MATH500 across the four RL checkpoints. Per-checkpoint full sets (n=578–593, with $\sim$15% subset variation) showed an apparent trajectory in the greedy/prob ratio; a controlled paired bootstrap on the 4-way intersection (n=501) does not reproduce it ($\Delta\text{ratio}_{\text{rl1900}-\text{base}} = +0.61\,[-2.67, +3.38]$, $P(\Delta{>}0)$=0.40; full table in Appendix P). The apparent trend is a problem-composition artifact; we recommend paired-set analysis as standard for cross-checkpoint patching comparisons at noise-floor effect sizes. RL-Zero leaves the in-domain head-level ghost structure unchanged.

Robustness to threshold choice. The two-peak structure and RL-Zero null are stable across a 5×5 grid of greedy/probability thresholds and four ratio cutoffs (Appendix N). The continuous GhostScore (Eq. 2) gives the same picture: L31.H13 tops the ranking at every instability peak (score=10.5 at stage1-354k, 9.0 at stage2-31k) and no RL checkpoint exceeds 1.5.

3.4. Cross-Architecture Replication

Ghost heads are not an OLMo artifact. We replicate the phenomenon on DeepSeek-Math-7B and Qwen2.5-Math-7B at the base checkpoint, using the same 200 3-digit addition problems and patching protocol as OLMo-3 (Table 4, base columns). Per-model n reflects how many problems pass each model's clean-corrupt validity filter (DeepSeek n=194, Qwen n=187, OLMo-3 n=194). All three architectures show a single dominant specialist head agreed upon by both metrics, alongside ghost heads with elevated greedy but suppressed distributional recovery.

The cross-architecture data also reveals *inverse ghost heads*: heads nearly invisible under greedy patching but carrying substantial distributional mass. On Qwen-Math, L3.H4 (greedy 0.046, dist 0.306) and L19.H24 (greedy 0.112, dist 0.339) carry more distributional recovery than any head visible in the top-5 greedy ranking except the shared specialist L0.H3. Greedy-only patching would miss these components entirely, underestimating the distributional impor-

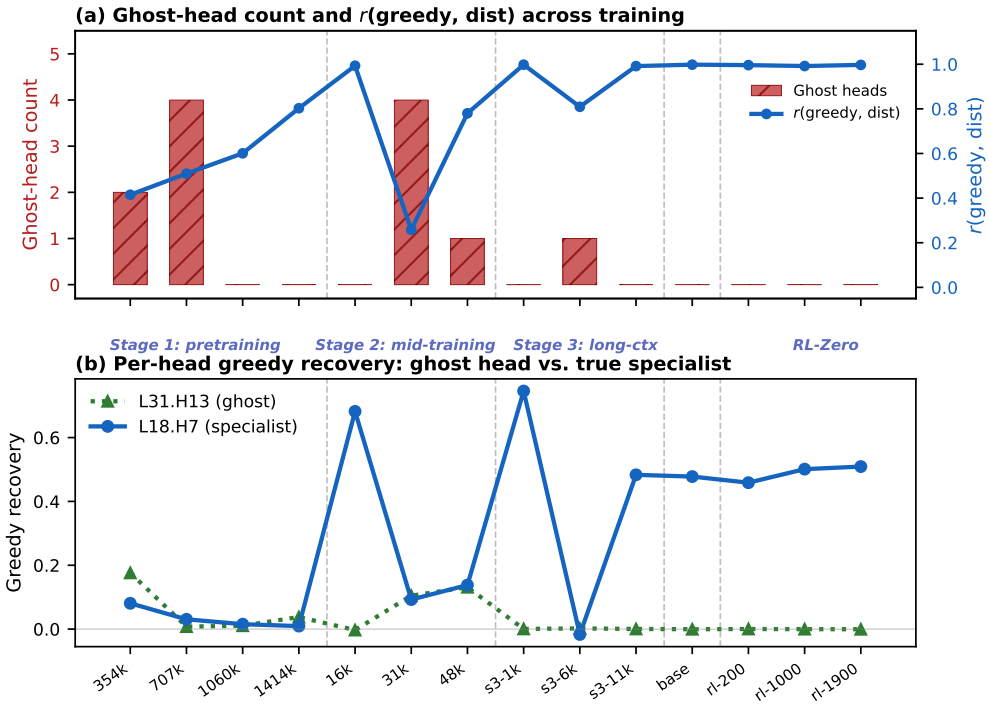

Figure 3. **Ghost-head trajectory across 14 checkpoints.** (a) Ghost-head count (bars, canonical criterion) and r(greedy, dist) (line). Ghost heads peak twice: in early pretraining and mid-stage mid-training. The correlation crashes to 0.26 at stage2-step31k, the worst point in the entire trajectory. (b) Greedy recovery of L31.H13 (ghost head, green) and L18.H7 (true specialist, blue). L31.H13 activates in early pretraining, dies during late pretraining, reactivates during mid-training, and is permanently eliminated by stage 3. Dashed vertical lines mark stage boundaries.

tance of mid-layer heads. On DeepSeek-Math, L27.H19 (greedy 0.011, dist 0.125) shows the same inverse pattern at $11\times$ ratio. Inverse ghost heads confirm that the greedy/distributional dissociation runs in both directions: some heads are overweighted by greedy patching (ghost heads) while others are underweighted (inverse ghost heads), and the full circuit picture requires both metrics. OLMo-2-7B replicates the temporal version of the phenomenon with an extreme anti-correlation ($r = -0.841$) at 25% through pretraining (Appendix F).

Reorganization under RL scales with weight-perturbation magnitude. Extending the analysis to the post-RL checkpoint of each model (Table 4, RL columns) reveals three qualitatively different reorganization signatures: OLMo-3 RL-Zero (0.05% weight change) leaves the picture essentially unchanged (r: .998 $\rightarrow$ .997, 83% ghost overlap); Qwen2.5-Math SimpleRL-Zero (0.16%) swaps the ghost set wholesale (6% overlap); DeepSeek-Math RL (3.0%, with an SFT stage before RL) produces the largest correlation crash (14% overlap), promoting its three base ghost heads to ordinary specialists rather than eliminating them. With n=3 architectures we do not claim a dose-response, but the r drop is monotone in $|\Delta W|$.

Table 4. **Cross-architecture ghost-head replication and reorganization under RL.** Base columns: HHI, r(g,d), n, top head, and an inverse-ghost example. RL columns: r(g,d) after RL, base$\rightarrow$RL change in correlation, percentile-defined ghost set overlap (% of base ghost heads still satisfying the per-model percentile criterion at RL), and relative weight change. Per-model percentile criterion: g above 95th percentile and ratio $g/(|p| + \epsilon)$ above the model's median plus 3 MAD. Same arithmetic patching protocol across all six runs.

	base				RL			
Model	HHI	r	n	Top	r	Δr	overlap	$\lvert\Delta W\rvert$
OLMo-3	.111	.998	194	L18.H7	.997	$-.001$	83%	.05%
Qwen-Math	.002	.427	187	L0.H3	.295	$-.132$	6%	.16%
DeepSeek-M.	.015	.444	194	L1.H7	.161	$-.283$	14%	3.0%

4. What Does RL-Zero Change?

Recent work argues that RL-Zero applies only tiny perturbations to the base model's weights (Yue et al., 2025; Zhao et al., 2025; Wu et al., 2025; Zhu et al., 2025). We test this with two independent probes and five weight-dependent measurements (Table 5). The two independent probes, noise fragility (where base and RL collapse equivalently at 0.01% MLP noise) and verifier reranking (where RL-trained models score base-generated reasoning within 1pp of the base model), both return null results on GSM8K, and both could

Table 5. **Seven MI tools on the RL-Zero transition.** Each returns a null or symmetric result on GSM8K math reasoning. Tools marked † are downstream of the same 0.05% relative weight change and are not statistically independent of one another. Only noise fragility and verifier reranking are independent of the weight change; the multi-tool null argument rests on these two.

Tool	Result	Ref.
Neuron overlap (top-50/MLP)†	∼100% overlap	Appx. A
Linear probing (carry, digit)†	identical to base	§3.2
Activation patching†	modest shifts	§3.2
Weight diffs + SVD†	full-rank, tiny	Appx. G
Transplant (n=500, 2 subsets)†	gap <2pp	Appx. B
Noise fragility (0.01% MLP)	base & RL equal	Appx. I
Verifier reranking (best-of-32)	RL ≈ base	Appx. J

in principle have shown a positive RL-Zero effect. Five additional measurements (neuron overlap, probing, activation patching, weight-diff SVD, transplant) are consistent but not independent: all are downstream of the same 0.05% relative weight change.

Weight-diff geometry. RL-Zero changes model weights by 0.05% relative to the base checkpoint (end of pretraining, before post-training); SFT accounts for 99.9% of the total pipeline weight change (Appendix G). All transitions produce full-rank perturbations, distinguishing RL-Zero from supervised distillation where rank-1 approximations suffice (Ward et al., 2025). When normalized by parameter count, MLP and attention are perturbed *equally* across four 7B pairs (Appendix H); the apparent MLP norm bias in absolute Frobenius norms is a parameter-count artifact. RL-Zero pairs perturb 10–60× less weight than RL-after-SFT or vanilla SFT pairs, suggesting tiny perturbations are specific to RL-Zero pipelines.

Cross-model head patching. To test whether the base-to-RL behavioral difference contains head-level ghost-head signatures, we run cross-model activation patching: for each attention head, we inject the base model's cached activation into the RL forward pass on the same arithmetic prompt and measure greedy and distributional recovery on the base-vs-RL output gap (n=200 problems, 1024 heads). The overall greedy/distributional correlation of the cross-model delta is $r = 0.69$, substantially lower than the within-model $r = 0.998$ at either endpoint, indicating more metric dissociation *between* models than *within* either model.

Under both canonical and ratio criteria, L31.H13, the canonical ghost head from the pretraining trajectory (§3.3), is the only head that qualifies as a cross-model ghost head (greedy 0.020, prob ≈ 0, ratio $6.5\times$). To control for the multiple-comparisons concern of selecting one head out of 1024, we run a permutation null (1000 iterations, per-example sign flip; details in Appendix O). The null distribution of the maximum GhostScore across the 1024 heads has 99th percentile 0.97 and maximum 1.25 over 1000 permutations. L31.H13's observed GhostScore of 2.02 is far above the null max ($p < 0.001$ Bonferroni-via-permutation), and L31.H13 is the top head by GhostScore in the observed data. The same head, exhibiting the same type of metric dissociation, is visible in both the training trajectory (where it peaks at every training-objective shift) and in the base-to-RL behavioral delta. This is the head-level continuity from pretraining ghost heads to the post-training argmax-layer account: L31.H13 is architecturally predisposed to the greedy/distributional dissociation because of its late-layer position and value-norm dynamics, and this predisposition manifests whenever the model's output distribution is perturbed. Cross-model effect sizes are roughly $10\times$ smaller than within-model effects, consistent with RL-Zero's 0.05% weight change producing proportionally weaker activation deltas.

The argmax interpretation. The convergent null across the two independent probes, together with the five additional weight-dependent tests, is consistent with the argmax/sampling-layer account of RL-Zero proposed from behavioral and geometric angles by Yue et al. (2025); Zhao et al. (2025); Wu et al. (2025); Zhu et al. (2025): RL-Zero does not reorganize identifiable head-level computation, and pass@1 and pass@32 diverge because RL-Zero is a whole-model distributional shift acting on the argmax-adjacent layer rather than on any individual circuit element. The seven-tools null is then *evidence for* the framing rather than a weakness it must excuse: standard MI tools find nothing to point at because there is no single-head locus to find.

5. Answer-Token Patching

We validate answer-token patching (§2) on 4-digit arithmetic, where it recovers L29.H5 (recovery 0.632), the same head identified by standard patching on that task. On GSM8K and MATH500, both tasks recruit a shared multi-step verification head L30.H22 (recovery 0.648 on GSM8K, 0.268 on MATH500), distinct from arithmetic's L29.H5: task-specific verification circuits that share a common multi-step verification head but diverge from the single-step arithmetic circuit. A control condition with a wrong chain-of-thought and a consistently wrong final answer produces diffuse recovery an order of magnitude below the main signal, ruling out mere consistency checking.

RL does not jointly improve verification and generation on any task. Figure 4 contrasts top-1 verification recovery (bootstrap 95% CI) against generation pass@1 (Wilson 95% CI, n=200) across stage 1, base, and RL. On arithmetic, pass@1 falls by 24pp from base to RL while verification recovery is unchanged (CIs fully overlapping).

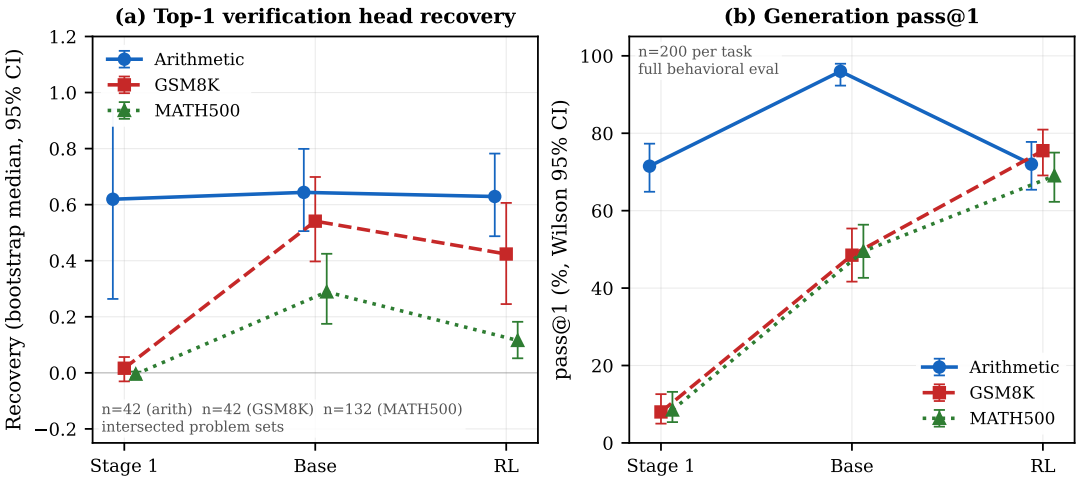

Figure 4. **Answer-token patching: verification top-1 head recovery vs. generation pass@1 across stage 1, base, and RL.** (a) Bootstrap median and 95% CI (n_{boot}=1000) of recovery for the top-1 verification head identified at base on each task: L29.H5 for arithmetic, L30.H22 for GSM8K and MATH500. Computed on the problem set intersected across all three checkpoints per task (n=42 arithmetic, n=42 GSM8K, n=132 MATH500). (b) pass@1 across the same checkpoints from the full behavioral evaluation (n=200 per task, T=0.8 sampling, Wilson 95% CI). Verification recovery emerges during pretraining; on arithmetic the base and RL CIs are indistinguishable; on GSM8K and MATH500 point estimates decline under RL with partially overlapping CIs.

On GSM8K and MATH500 the behavioral movement is opposite (pass@1 rises by 27/20pp), but verification recovery is no higher under RL. RL produces no verification gain on either the task where pass@1 falls or the tasks where it rises.

6. Discussion

Implications for self-verification. The convergent verification null across answer-token patching (§5) and verifier reranking (Appendix J, base and RL within 1pp on GSM8K) provides a mechanistic argument for explicitly trained process reward models (Lightman et al., 2024): outcome-level RL does not produce useful self-verifiers as a side effect.

Limitations. Our temporal analysis covers a single model family (OLMo) with single-checkpoint cross-architecture snapshots. The ghost-head criterion is threshold-sensitive (qualitative pattern robust, Appendix N) and not derived from a noise model. The verifier-reranking probe (n=100, 10pp variance) is the only weight-independent null test; a larger run would strengthen it. Answer-token patching is validated only on arithmetic. We do not test component interactions or scaling across model sizes.

Future work. Tracking ghost heads across other mid-training pipelines would test whether the recurrent instability is general or specific to OLMo-3's data mixture. A targeted value-projection scaling experiment at the ghosting checkpoint would isolate the value-output contribution more cleanly than activation patching. The architectural specificity of ghost heads predicts that architectures without softmax attention (e.g., linear attention, state-space models) should not exhibit the dissociation.

7. Related Work

Temporal circuit tracking. Tigges et al. (2024) (Pythia, syntax), Nikankin et al. (2025) (Pythia, arithmetic), and Hakimi et al. (2025) (OLMo-7B, factual recall) track circuits across pretraining; we extend temporal analysis to the full pretraining-through-RL pipeline with dual metrics.

Dormant heads and attention sinks. Ghost heads build on work showing attention magnitude is a poor proxy for causal importance (Kobayashi et al., 2020; Xiao et al., 2024; Guo et al., 2024; Sandoval-Segura et al., 2026); we extend the single-checkpoint metric-dissociation of Zhang & Nanda (2024) temporally and show dormancy criteria miss ghost heads at the stage they cause the dissociation.

RL-Zero mechanism. Unlike rank-1 SFT distillation (Ward et al., 2025), RL-Zero diffs are full-rank and 10–60× smaller. Park et al. (2025) find GRPO creates small targeted head sets; our null is consistent if those changes scale with $|\Delta W|$ (Appendix R). Our pass@32-preserved finding is one data point for the bounded-argmax view (Yue et al., 2025; Wen et al., 2025; Venhoff et al., 2026).

Acknowledgements

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

References

Chan, L., Garriga-Alonso, A., Goldowsky-Dill, N., Greenblatt, R., Nitishinskaya, J., Radhakrishnan, A., Shlegeris, B., and Thomas, N. Causal Scrubbing: A method for rigorously testing interpretability hypotheses. AI Alignment Forum, December 2022. URL https://www.alignmentforum.org/posts/JvZhhzycHu2Yd57RN/causal-scrubbing-a-method-for-rigorously-testing.

Geiger, A., Ibeling, D., Zur, A., Chaudhary, M., Chauhan, S., Huang, J., Arora, A., Wu, Z., Goodman, N., Potts, C., and Icard, T. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025. URL https://jmlr.org/papers/v26/23-0058.html.

Guo, T., Pai, D., Bai, Y., Jiao, J., Jordan, M. I., and Mei, S. Active-dormant attention heads: Mechanistically demystifying extreme-token phenomena in LLMs. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2410.13835.

Hakimi, A. D., Modarressi, A., Wicke, P., and Schütze, H. Time course MechInterp: Analyzing the evolution of components and knowledge in large language models, 2025. URL https://arxiv.org/abs/2506.03434.

Heimersheim, S. and Nanda, N. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*, 2024.

Kobayashi, G., Kuribayashi, T., Yokoi, S., and Inui, K. Attention is not only a weight: Analyzing Transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7057–7075, Online, nov 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.574. URL https://aclanthology.org/2020.emnlp-main.574/.

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 (ICLR)*, 2024.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 35, 2022.

Nikankin, Y., Reusch, A., Mueller, A., and Belinkov, Y. Arithmetic without algorithms: Language models solve Math with a bag of heuristics. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.21272.

OLMo Team. OLMo 3, 2025. URL https://arxiv.org/abs/2512.13961. Model release; allenai/Olmo-3-1025-7B on HuggingFace.

OLMo Team, Walsh, P., Soldaini, L., Groeneveld, D., Lo, K., Arora, S., Bhagia, A., Gu, Y., Huang, S., Jordan, M., Lambert, N., Schwenk, D., Tafjord, O., Anderson, T., Atkinson, D., Brahman, F., Clark, C., Dasigi, P., Dziri, N., Guerquin, M., Ivison, H., Koh, P. W., Liu, J., Malik, S., Merrill, W., Miranda, L. J. V., Morrison, J., Murray, T., Nam, C., Pyatkin, V., Rangapur, A., Schmitz, M., Skjonsberg, S., Wadden, D., Wilhelm, C., Wilson, M., Zettlemoyer, L., Farhadi, A., Smith, N. A., and Hajishirzi, H. 2 OLMo 2 furious. In *Conference on Language Modeling (COLM)*, 2025.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Johnston, S., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Park, Y., Jeong, M., and Kang, J. Thinking Sparks!: Emergent attention heads in reasoning models during post training. *arXiv preprint arXiv:2509.25758*, 2025.

Sandoval-Segura, P., Wang, X., Panda, A., Goldblum, M., Basri, R., Goldstein, T., and Jacobs, D. Identifying and evaluating inactive heads in pretrained LLMs. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2504.03889.

Tigges, C., Hanna, M., Yu, Q., and Biderman, S. LLM circuit analyses are consistent across training and scale. In *Advances in Neural Information Processing Systems*, 2024.

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 (ICML)*, 2026. Spotlight.

Wang, K. R., Variengien, A., Conmy, A., Shlegeris, B., and Steinhardt, J. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2211.00593.

Ward, J., Riechers, P. M., and Shai, A. Rank-1 LoRAs encode interpretable reasoning signals. In *NeurIPS 2025 Workshop on Mechanistic Interpretability*, 2025. URL https://openreview.net/forum?id=l4ifwCE3uw.

Wen, X., Liu, Z., Zheng, S., Ye, S., Wu, Z., Wang, Y., Xu, Z., Liang, X., Li, J., Miao, Z., Bian, J., and Yang, M. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base LLMs. *arXiv preprint arXiv:2506.14245*, 2025.

Wu, F., Xuan, W., Lu, X., Liu, M., Dong, Y., Harchaoui, Z., and Choi, Y. The invisible leash: Why RLVR may or may not escape its origin. *arXiv preprint arXiv:2507.14843*, 2025.

Xiao, G., Tian, Y., Chen, B., Han, S., and Lewis, M. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2309.17453.

Yue, Y., Chen, Z., Lu, R., Zhao, A., Wang, Z., Yue, Y., 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 (NeurIPS)*, 2025.

Zhang, F. and Nanda, N. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, 2024.

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

Zhu, H., Zhang, Z., Huang, H., Su, D., Liu, Z., Zhao, J., Fedorov, I., Pirsiavash, H., Sha, Z., Lee, J., Pan, D. Z., Wang, Z., Tian, Y., and Tai, K. S. The path not taken: RLVR provably learns off the principals. *arXiv preprint arXiv:2511.08567*, 2025.

A. Full Recovery Tables

Bootstrap confidence intervals (trimmed means, 10,000 BCa resamples) for key attention heads across all OLMo-3 checkpoints are provided in the code and data release. Key results at base: L18.H7 trimmed probability recovery is 0.488 [0.392, 0.580]; L31.H13 (the ghost head) is approximately zero (-0.0002 [-0.0006, 0.0003]), consistent with its elimination by stage 3.

B. Weight Transplant Details

Mean accuracy across 3 repeats per condition; per-repeat values are in the released CSV. **RL transplant** (GSM8K, n=500 per subset, T=0.8): Base 43.9%/43.2%, RL-Zero 44.1%/44.7%, Base MLP + RL Attn 45.3%/44.3%, RL MLP + Base Attn 44.8%/42.6% (subset 1 / subset 2). Gaps of 0.2pp and 1.5pp, both within per-repeat noise; the hybrid direction is not consistent between subsets. An earlier n=100 pilot had a 6pp gap that did not survive replication. **Mid-training transplant** (3 repeats): on arithmetic, Stage1 $\sim$66%, Base $\sim$89%, Stage1 MLP+Base Attn $\sim$52%, Base MLP+Stage1 Attn $\sim$74% — base MLPs preserve most performance, base attention does not. On GSM8K and MATH500 every condition collapses to $\leq$10%, ruling out a clean decomposition. **Arithmetic RL transplant**: 1.4pp gap at ceiling, no signal.

C. Multi-Step Arithmetic

HHI by step count, with 95% percentile bootstrap intervals (2000 resamples over problems, computed from per-(problem, layer, head) mean recoveries): 1-step 0.056 $[0.034, 0.069]$ ($n = 250$), 2-step 0.037 $[0.020, 0.047]$ ($n = 200$), 3-step 0.033 $[0.019, 0.040]$ ($n = 150$), 4-step 0.038 $[0.021, 0.048]$ ($n = 100$). The 1-step value is significantly higher than each of the multi-step values (CIs barely overlap), but 2-, 3-, and 4-step are statistically indistinguishable from each other. The clean signal is therefore single-step versus multi-step (a $\sim$33% drop in concentration), not a smooth monotonic gradient with step count. L18.H7 remains the top head across all step counts. mlp_22 dominates multi-step computation (recovery 0.398–0.470) rather than mlp_30/31. Subsolution probing: intermediate results are visible at L31 for 1–2 step problems (rank 224–348) but degrade for 3–4 steps (rank 1038–4064).

D. GSM8K and MATH500 Direct Patching

Direct single-token patching on GSM8K and MATH500 produces weak signal (36–44% of examples have clean logit $>$ corrupt logit) due to the denominator problem: single-number corruption does not strongly affect first-token prediction for multi-step tasks. MLP dominance holds

at the layer level (mlp_31 recovery 0.298–0.390 across checkpoints). Head-level concentration gradient: HHI decreases from 0.111 (arithmetic) to 0.026 (GSM8K) to 0.015 (MATH500). These results should be interpreted with caution given the signal quality limitations.

E. CoT Comparison Details

Base model: pass@1 = 52%, pass@32 = 96% on 50 GSM8K problems. RL model: pass@1 = 50%. Per-token log-probabilities: base correct chains mean -0.294; RL chains mean -0.310. RL generates longer chains (271 versus 223 tokens) with more computation steps (3.92 versus 3.77).

F. OLMo-2 Cross-Model Replication

We replicate the temporal ghost-head phenomenon on OLMo-2-7B over 4 checkpoints spanning pretraining and mid-training (original release; 8 pretraining + 3 mid-training checkpoints evaluated for behavioral results, 4 used for full head-level patching). At 25% through pretraining, OLMo-2 shows an extreme anti-correlation between greedy and distributional head recovery ($r = -0.841$), more severe than any OLMo-3 checkpoint. L19.H22 is the single dominant head throughout; greedy 0.248 and probability recovery -0.775 at step 228k, converging to greedy 0.184 / prob 0.183 ($r = 0.725$) at the final base model. The qualitative phenomenon (metric convergence during training) replicates; the quantitative magnitude is architecture-specific.

G. Weight-Diff Geometry

Table 6. **Weight changes across the OLMo-3 post-training pipeline.** SFT accounts for 99.9% of total weight change; DPO and RLVR contributions are two to three orders of magnitude smaller. All diffs are full-rank.

Transition	$\|\Delta W\|_{\text{attn}}$	$\|\Delta W\|_{\text{MLP}}$	Relative
Base → SFT	219.2	365.0	25.0%
SFT → DPO	0.74	0.95	0.06%
DPO → RLVR	1.38	2.04	0.14%
Base → RL-Zero	0.47	0.72	0.05%

SVD decomposition: all transitions produce full-rank perturbations (top-10 singular values capture less than 3% of total), distinguishing RL-Zero from supervised distillation where rank-1 approximations suffice (Ward et al., 2025).

H. Cross-Architecture Weight Diffs

The two RL-Zero pairs perturb 10–60× less of the model weight in relative terms than either DeepSeek-Math RL (which followed an SFT stage) or vanilla Qwen-Instruct SFT, suggesting that tiny perturbations are a property of RL-Zero pipelines specifically, not of RL in general. Vanilla SFT (Qwen-Instruct) is also full-rank in our analysis, consistent with the view that the rank-1 approximations of Ward et al. (2025) are specific to *distillation* SFT rather than supervised fine-tuning in general.

Table 7. **Four 7B pairs across two architectures and two post-training paradigms.** M/A ratios are reported both as absolute Frobenius norms (abs) and as relative changes normalized by base-weight norm (rel). After normalization MLP and attention are perturbed equally; the apparent MLP bias in absolute norms is a parameter-count artifact. SV = singular value, eff. rank = effective rank of ΔW.

Pair	Type	M/A (abs)	M/A (rel)	Rel. Δ	Top-10 SV	Eff. rank
OLMo-3 RL-Zero	RL-Zero	1.52×	1.00×	.05%	1.5–1.9%	full
Qwen SimpleRL	RL-Zero	2.57×	1.00×	.16%	3.3%	2779
DeepSeek-Math RL	RL→SFT	1.47×	1.02×	3.00%	1.7%	3374
Qwen-2.5 Instruct	SFT	2.66×	1.04×	1.66%	1.5%	3065

I. Noise Fragility

We add isotropic Gaussian noise to MLP weights at scales of 0.01–1% of each parameter matrix's Frobenius norm and measure pass@1. At 0.01% noise, both RL and base collapse on GSM8K (RL: $48\% \to 2\%$; base: $53\% \to 2\%$), and arithmetic collapses from 91% to 0%. The fragility is universal and is a general property of transformer MLP weights, not a signature of RL's specific perturbations.

J. Verifier Reranking

We generate 32 samples from the base model on n=100 GSM8K problems, score each sample by mean length-normalized log-probability under either the base or the RL model, and select the highest-scoring sample. Two independent runs with different random seeds give:

- Run 1 (max 256 tokens): Base pass@1 57%, pass@32 (oracle) 97%, Base-gen+Base-verify 65%, Base-gen+RL-verify 64%.
- Run 2 (max 1024 tokens): Base pass@1 57%, pass@32 99%, Base-gen+Base-verify 54%, Base-gen+RL-verify 55%.

The absolute reranked accuracy swings 10pp between runs (high variance of best-of-32 on n=100), but in both runs the two verifiers are within 1pp of each other. Log-probability self-scoring recovers at most a small fraction of the pass@1 → pass@32 gap, and the choice of reranker (base vs. RL-Zero) does not matter within noise in either run.

K. Answer-Token Patching: Example Prompts

Each clean/corrupted prompt pair ends with the verification cue " `The solution is`"; patching is performed on

the next token, where the model commits to "correct" or "incorrect."

Arithmetic (4-digit).

- Clean: `Q: What is 4729 + 1583?\nA: The answer is 6312. The solution is`
- Corrupt: `Q: What is 4729 + 1583?\nA: The answer is 7414. The solution is`

Only the final answer is altered; the operands are unchanged.

GSM8K.

- Clean: `Q: {question}\nA: {chain-of-thought} The answer is 18. The solution is`
- Corrupt: `Q: {question}\nA: {chain-of-thought} The answer is 42. The solution is`

The chain of thought is preserved verbatim; only the final numeric answer changes.

Control (wrong CoT, consistent wrong answer). The control replaces one numeric step inside the chain of thought *and* updates the final answer to match, so the corrupt prompt is internally consistent. Recovery on this control is an order of magnitude smaller than on the main condition (§5), ruling out mere coherence tracking.

L. Per-Category MATH500 Verification

MATH500 is organized into 7 MATH subject categories. We compute per-category sum of positive head recoveries under answer-token patching. Between base and RL-Zero-step1900, Algebra drops from 11.54 to 2.59 (-78%) and Intermediate Algebra from 12.42 to 6.53 (-47%), while Counting & Probability, Number Theory, Prealgebra, and Geometry are stable or grow. The collapsed categories are precisely those that had previously used L29.H5 / L30.H22, the same specialist heads as arithmetic and GSM8K, suggesting RL specializes these heads further on arithmetic-shaped problems at the expense of the broader Algebra distribution. MATH500 pass@32 at the task level remains preserved ($89\% \rightarrow 92\%$), so the category-level contraction does not aggregate to a whole-task collapse.

M. Verification-Signal Aggregate Trajectory

Beyond the top-1 head trajectory reported in Figure 4, we also report the aggregate sum of positive head recoveries across the entire 32-layer attention stack from answer-token patching, computed on the problem set intersected across stage 1, base, and RL per task (Table 8). Arithmetic aggregate signal contracts from base to RL; MATH500 and GSM8K are roughly flat at the aggregate level (Table 8), though the top-1 verification head (L30.H22) declines on both tasks (GSM8K: $.541 \rightarrow .424$; MATH500: $.289 \rightarrow .116$; Figure 4). We do not claim within-task significance for the aggregate contraction.

Table 8. **Sum of positive head recoveries from answer-token patching across stage 1, base, and RL.** Computed on the intersected problem set per task (n=42 arithmetic, n=42 GSM8K, n=132 MATH500). Bootstrap 95% CIs (1000 resamples). Arithmetic contracts from base to RL; GSM8K and MATH500 are roughly flat at the aggregate level, though the top-1 head (L30.H22) declines on both.

Task	stage 1	base	RL
Arithmetic (n=42)	5.5 [4.0, 8.8]	4.8 [3.8, 6.5]	3.1 [2.6, 3.9]
GSM8K (n=42)	4.2 [2.3, 7.8]	4.2 [2.8, 6.0]	3.3 [2.4, 4.7]
MATH500 (n=132)	12.3 [10.7, 14.1]	13.1 [11.6, 14.7]	13.3 [11.7, 15.0]

Per-category analysis of MATH500 (Appendix L) shows the largest point-estimate drops occur in Algebra and Intermediate Algebra, the categories that share specialist heads with arithmetic, but the same caveat about intersected sample size and bootstrap CIs applies.

N. Ghost-Head Threshold Sensitivity

To rule out that our findings depend on the canonical cutoff (greedy > 0.05, $|\text{prob}| < 0.02$), we sweep greedy $\in \{0.02, 0.05, 0.08, 0.10, 0.15\} \times |\text{prob}| \in \{0.005, 0.01, 0.02, 0.03, 0.05\}$ across all 14 checkpoints, and apply two alternative criteria: a ratio criterion and the continuous GhostScore (Eq. 2). Table 9 reports four representative settings.

Ratio criterion ($g/(|p| + \epsilon)$). At ratio > 5: stage1-354k yields 3, stage1-707k 2, stage2-31k 2, every RL-Zero checkpoint 0 — reproducing the two-peak structure with no absolute threshold. At ratio > 10, only stage1-354k remains (1 head, L31.H13).

Continuous GhostScore. Ranking all 1024 heads by GhostScore (Table 10): L31.H13 is top-1 at all three instability peaks (scores 6.7–10.5); elsewhere the top head has score < 2, and no RL-Zero checkpoint reaches 1.5.

All sweep outputs are reproduced by the script `ghost_threshold_sensitivity.py` and released alongside the code.

Table 9. **Canonical ghost-head count is robust across thresholds.** Ghost counts at four representative (g, p) settings. The two-peak structure (stage1-354k/707k, stage2-31k) and the RL null are stable across every combination tested.

Checkpoint	g>.02 p<.02	g>.05 p<.02	g>.08 p<.02	g>.10 p<.02
stage1-354k	15	**2**	1	1
stage1-707k	24	**4**	0	0
stage1-1060k	14	0	0	0
stage1-1414k	10	0	0	0
stage2-16k	3	0	0	0
stage2-31k	14	**4**	1	1
stage2-48k	2	1	1	1
stage3-1k	3	0	0	0
stage3-6k	6	1	0	0
stage3-11k	2	0	0	0
base	1	0	0	0
rlzero-200	0	0	0	0
rlzero-1000	0	0	0	0
rlzero-1900	3	0	0	0

O. Cross-Model Head Patching Details

For each (layer, head) across 32 layers $\times$ 32 heads, we inject the base model's cached pre-o_{proj} activation into the RL forward pass at the final token position on 200 arithmetic problems. Recovery is $(\ell_{\text{patched}} - \ell_{\text{RL}})/(\ell_{\text{base}} - \ell_{\text{RL}})$ for greedy and analogously for log-probability. Validity filter $|\ell_{\text{base}} - \ell_{\text{RL}}| > 0.1$ keeps all 200 problems. L31.H13 is the only head satisfying both the canonical (greedy > 0.01, $|p| < 0.005$) and ratio ($> 3\times$) criteria.

Permutation null over 1024 heads. 1000 iterations of per-example sign flip (± 1, $p = 0.5$); per iteration we recompute trimmed-mean recoveries per head and record the maximum GhostScore across all 1024 heads. This preserves within-example difficulty structure while destroying the consistent direction of effect. The null max-GhostScore distribution has mean 0.32, median 0.27, $p_{95} = 0.69$, $p_{99} = 0.97$, maximum 1.25. L31.H13's observed GhostScore 2.02 exceeds the null max, giving $p < 0.001$ Bonferroni-corrected. L31.H13 is rank 1 of 1024 by GhostScore (4th by greedy alone).

P. MATH500 L29.H5 Paired-Bootstrap Non-Result

We probed whether L29.H5 develops an RL-induced ghost signature on MATH500 (§3.3). Two analyses give conflicting answers; the paired analysis is the controlled comparison.

Per-checkpoint full sets (uncontrolled, $n\approx580$, $\sim$15% subset variation across checkpoints):

Table 10. **Top head by continuous GhostScore at each checkpoint.** L31.H13 dominates at the three instability moments (scores 6.7–10.5); everywhere else the top head has GhostScore close to 1, indistinguishable from noise.

Checkpoint	Top-1 head	GhostScore
stage1-354k	**L31.H13**	**10.5**
stage1-707k	L26.H25	$\sim$1.2
stage1-1060k	L19.H21	1.4
stage1-1414k	L24.H13	1.3
stage2-16k	L18.H7	1.0
stage2-31k	**L31.H13**	**9.0**
stage2-48k	**L31.H13**	**6.7**
stage3-1k	L29.H5	2.6
stage3-6k	L29.H5	4.3
stage3-11k	L15.H10	1.8
base	L12.H28	1.2
rlzero-200	L18.H7	1.1
rlzero-1000	L24.H19	1.3
rlzero-1900	L16.H2	1.5

Table 11. **L29.H5 on MATH500 – per-checkpoint full sets (uncontrolled).** Each checkpoint uses its own problem subset; subsets differ by $\sim$15% across checkpoints. The apparent non-monotone trajectory in the ratio column is a problem-composition artifact (see Table 12).

Checkpoint	n	g	p	ratio
base	578	.0228	.0028	2.93
rlzero-step200	579	.0177	.0027	2.29
rlzero-step1000	584	.0348	.0057	3.27
rlzero-step1900	593	.0513	.0070	4.28

4-way intersected paired set (controlled, n=501, 10 000-iteration percentile bootstrap 95% CIs):

Table 12. **L29.H5 on MATH500 – 4-way intersected paired set with bootstrap CIs.** Same 501 MATH500 problems at every checkpoint; 10 000-iteration percentile bootstrap 95% CIs in brackets. The non-monotone trajectory of the unrestricted analysis disappears: ratios are not monotone across the four checkpoints, and all CIs overlap heavily.

Checkpoint	g [95% CI]	p [95% CI]	ratio [95% CI]
base	.021 [.011, .032]	.002 [-.005, .010]	2.87 [1.15, 4.84]
rlzero-step200	.019 [.008, .029]	.002 [-.006, .011]	2.57 [0.82, 4.47]
rlzero-step1000	.025 [.012, .040]	.007 [-.007, .020]	2.15 [0.77, 5.34]
rlzero-step1900	.031 [.014, .049]	.009 [-.005, .024]	2.16 [0.77, 6.23]

Paired difference (rlzero-step1900 − base, same 501 problems): $\Delta g = +0.005\,[-0.010, +0.020]$, $\Delta p = -0.003\,[-0.012, +0.007]$, $\Delta\text{ratio} = +0.61\,[-2.67, +3.38]$; $P(\Delta\text{ratio}>0) = 0.395$, direction not favored.

Methodological note. The discrepancy is a problem-composition artifact: with $\sim$15% subset variation and noise-floor effect sizes (greedy in $[.018, .051]$, prob in $[.003, .009]$), per-checkpoint composition differences manu-

facture the apparent monotone trajectory of Table 11. We do not interpret L29.H5 as RL-induced ghost-head evidence on MATH500, and recommend paired-set analysis with bootstrap CIs as standard for cross-checkpoint comparisons at noise-floor effects.

Q. Value-Norm vs. GhostScore Across Checkpoints

The mechanism story in §3.2 predicts that ghost heads arise because softmax routing decouples a head's contribution to the residual stream from its causal effect on the argmax: a head whose value-state output has been driven down can still shift the argmax through its attention pattern. This predicts that ghost heads should have *elevated*, not collapsed, value norms at the moment they are causing the dissociation. We test this on the six OLMo-3 checkpoints for which we have both per-example value-state norms ($\|V \cdot x\|_2$ at the final token, computed on the same 200 arithmetic problems used for patching) and dual-metric head patching data.

Table 13. **L31.H13 value-norm and GhostScore trajectory.** The canonical pretraining ghost head re-elevates its value norm during the mid-training instability (stage2-step48k: 5.95, GhostScore 6.7) before collapsing to dormancy at base (0.74, GhostScore 0) and remaining there through RL. The per-checkpoint Spearman correlation between mean value norm and GhostScore across all (layer, head) pairs is positive at every checkpoint (column ρ). The ratio column gives mean value-norm of canonical-criterion ghost heads ($g>0.05$, $|p|<0.02$) divided by mean of the rest; "—" marks checkpoints with zero ghost heads under that criterion.

Checkpoint	L31.H13 vnorm	L31.H13 GS	ρ	ratio
stage1-step354k	**13.96**	**12.20**	+0.01	**5.93**×
stage1-step1414k	1.47	3.78	+0.15	—
stage2-step48k	**5.95**	**6.71**	+0.29	**3.94**×
base	0.74	0.00	+0.18	—
rlzero-step200	0.74	0.03	+0.37	—
rlzero-step1900	0.71	0.00	+0.25	—

Three observations: (i) L31.H13's value norm peaks twice (stage1-354k, stage2-step48k) in lockstep with its GhostScore peaks. (ii) Value norm and GhostScore are positively correlated at every checkpoint (Spearman $\rho \in [+0.01, +0.37]$); the correlation is weakest at stage1-354k where L31.H13 dominates the distribution. (iii) At base and RL, L31.H13's value norm has collapsed to $\sim$0.7, an order of magnitude below L18.H7 (8.37 at base), at which point output-norm dormancy criteria correctly identify it as inactive — but only after the dissociation it caused has resolved.

Coverage note: stage2-step48k and rlzero-step200 use 4-layer head-patching CSVs (layers 18, 29, 30, 31), so the Spearman ρ at those checkpoints is over $n = 128$ heads; the L31.H13 entry is unaffected.

R. Reconciliation with Thinking Sparks!

Park et al. (2025) track emergent attention heads during GRPO post-training and find that GRPO operates in a "dynamic search mode," iteratively activating, evaluating, and pruning a small, targeted set of heads, while SFT adds stable heads cumulatively. This is in apparent tension with our null results on RL-Zero (96–100% neuron overlap, no detectable transplant signal). Two compatible readings. First, Park et al. report that GRPO's final set of emergent heads is "small and targeted"; if the targeted changes on OLMo-3 RL-Zero are below the aggregate-measurement threshold of our neuron overlap and per-head transplant analyses, we would observe exactly the null results we report. Second, OLMo-3 RL-Zero may sit at the quiet end of the RL-dynamics spectrum: our weight diffs are 10–60× smaller than those of RL-after-SFT pipelines (Appendix H), so the same mechanistic phenomena may produce detectable signals in their settings while staying below detection in ours. Our findings and theirs are consistent if GRPO-induced head changes are narrow, targeted, and scale with the magnitude of the weight change.