Ghost Heads Across Training

When greedy and distributional activation patching disagree

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

Tracked arithmetic, GSM8K and MATH500 circuits across **14 OLMo-3-7B checkpoints** — pretraining, mid-training, long-context, RL-Zero — patching every head with **two metrics at once**: greedy logit recovery g and distributional log-prob recovery p.

WHAT WE FOUND

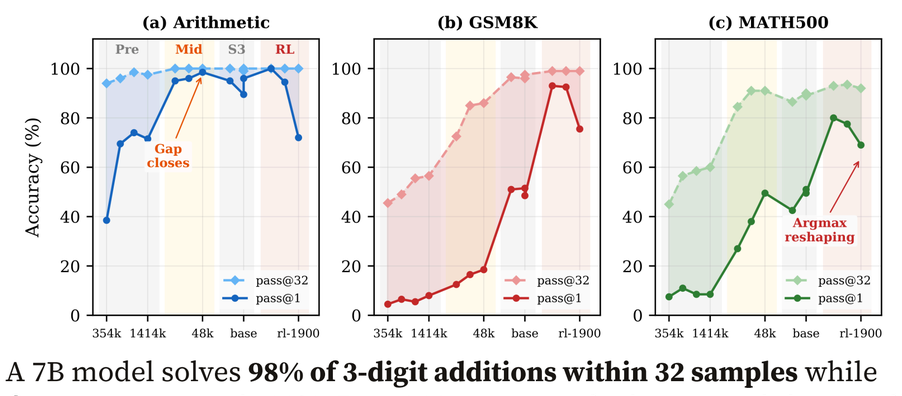
Ghost heads: heads with high g but ≈0 p. They spike at every training-objective shift, decay within phases, and are gone by the base model. Replicates in 3 model families. RL-Zero adds none — we detect **no head-level reorganization**.

WHY IT MATTERS

Single-metric patching can manufacture circuit stories. The apparent "relay" of dominant heads across pretraining is a metric artifact. Report both metrics; $r(g,p)$ doubles as a free circuit-stability diagnostic during training.

01 The reliability gap

MOTIVATION



A 7B model solves **98% of 3-digit additions within 32 samples** while first-try accuracy is 38%. Pass@32 saturates during **pretraining** on all three tasks; pass@1 lags by 25–50pp. **Mid-training** closes the gap. **RL-Zero** briefly peaks pass@1, then degrades it — with pass@32 preserved throughout, ruling out capability destruction.

PASS@1 / @32	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

Each task recruits its own specialist head — `L18.H7` arithmetic, `L26.H31` GSM8K, `L20.H21` MATH500 — on a shared MLP substrate (`mlp_31` top everywhere). *What is the mechanistic substrate of the gap, and what does each training stage change?*

MORE lab ICT for good · Deusto

MORElab · DeustoTech

02 Ghost heads: one patch, two verdicts

CORE FINDING

DEFINITION

A **ghost head** flips the winning token back to correct when patched (high greedy g) but doesn't shift probability mass onto it ($p \approx 0$) — it changes which answer wins, not how much the model believes it. `GhostScore(h) = max(0,g) / max(|p|+ε, φ)`

19:1

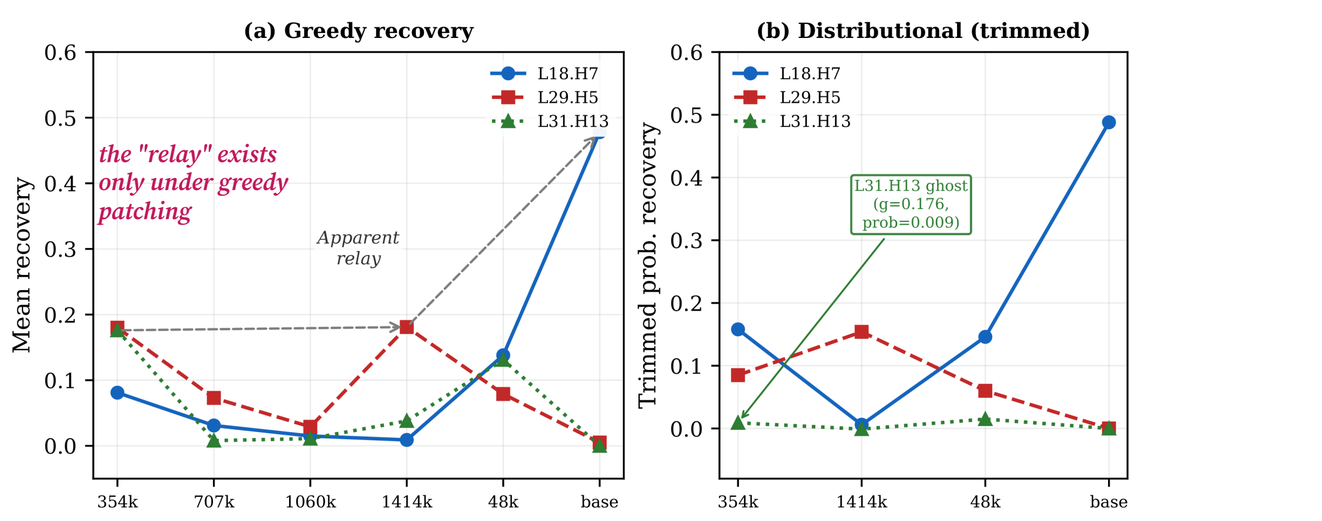
causal dissociation of `L31.H13` at step 354k — restoring it shifts the argmax 17.6% of the way back; probability mass moves 0.9%

5.9×

value-state norm vs. model average — ghosts are **strong but misdirected**, not dormant; dormancy criteria fire only at base, after the dissociation has resolved

0 MLPs

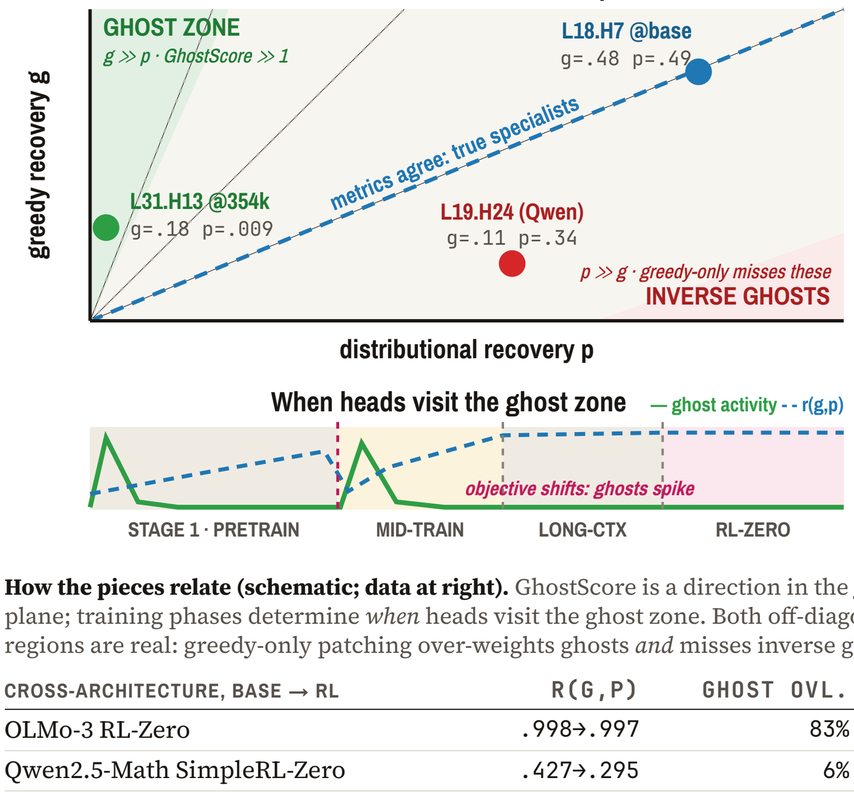
ghost signatures at any checkpoint — softmax forces every *head* to route mass regardless of task alignment; MLP neurons have no such constraint

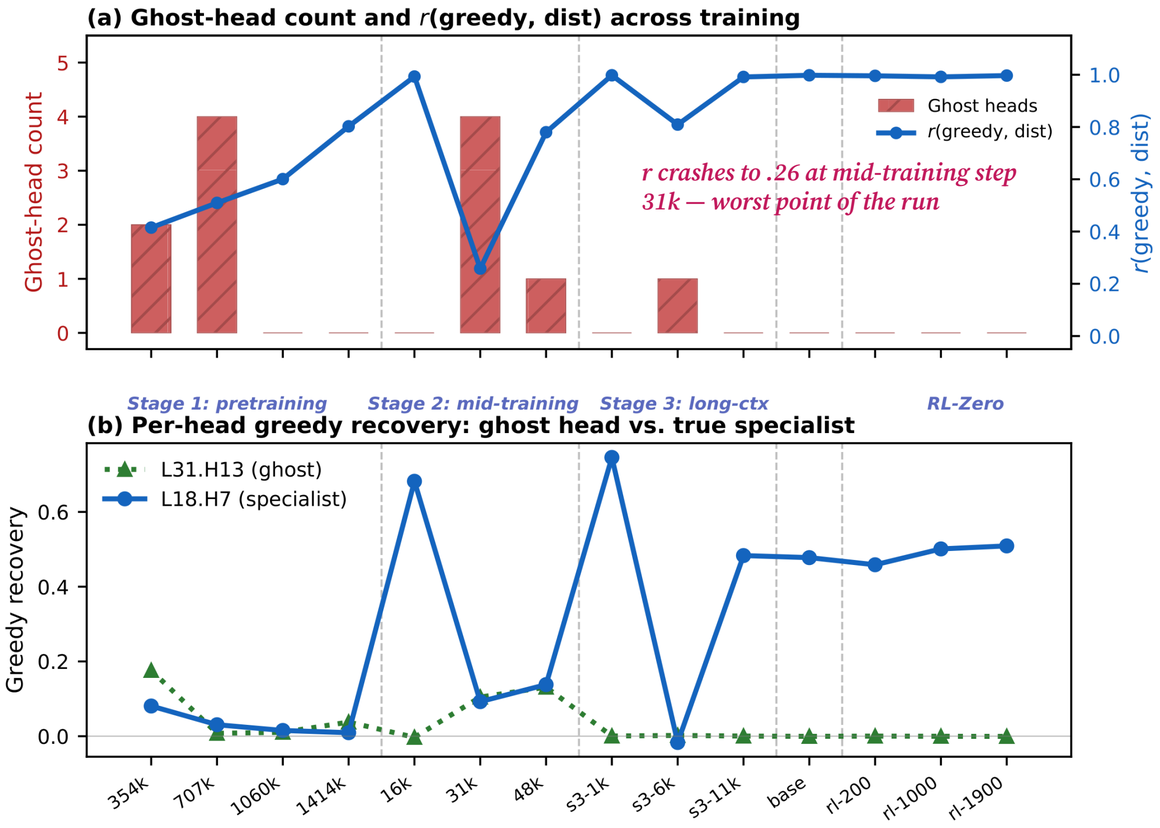


Greedy patching (left): three heads relay dominance across pretraining, `L31.H13` → `L29.H5` → `L18.H7`**. Distributional patching (right) dissolves it:** `L18.H7` dominates throughout; `L31.H13` sits at $p \approx 0$ for its entire trajectory. The transfer story is an artifact of the metric. Value norm correlates with GhostScore at every checkpoint (Spearman ρ +.01 to +.37) — consistent with a routing-vs-contribution mechanism; the causal value-scaling test is future work.

03 One map: two metrics, GhostScore, and the training phases

14-CHECKPOINT TRAJECTORY · ARITHMETIC





How the pieces relate (schematic; data at right). GhostScore is a direction in the g–p plane; training phases determine *when* heads visit the ghost zone. Both off-diagonal regions are real: greedy-only patching over-weights ghosts *and* misses inverse ghosts.

CROSS-ARCHITECTURE, BASE → RL	R(G,P)	GHOST OVL.	|ΔW|
OLMo-3 RL-Zero	.998→.997	83%	0.05%
Qwen2.5-Math SimpleRL-Zero	.427→.295	6%	0.16%
DeepSeek-Math RL (SFT first)	.444→.161	14%	3.0%

Replicates at base in all three architectures; OLMo-2 shows the temporal version (r = −0.84 at 25% of pretraining). Post-RL correlation drop is monotone in |ΔW| (n=3 — no dose-response claim).

Ghost heads are a recurrent training instability, not a monotone artifact. Two peaks (4 heads each) at the two objective shifts: early pretraining and the switch to curated mid-training data. Mid-training is **oscillatory** — form, disrupt, reform within 48k steps — then it resolves: at base, $r(g,p)$ = .998 and `L18.H7` is confirmed by both metrics (`g=.478, p=.488`).

04 What does RL-Zero change?

NO HEAD-LEVEL LOCUS

Almost nothing we can pin down. RL-Zero barely moves the weights — **0.05%** (SFT does 99.9% of the post-training pipeline). A battery of interpretability tools — neuron overlap, probing, patching, transplants, noise tests — turn up **no head-level reorganization** on in-domain math.

The only trace is in metric space, not the weights: `L31.H13` — the same ghost from pretraining — is the lone head the base→RL comparison flags.

Consistent with a **whole-model argmax/sampling shift** (Yue, Zhao, Wu, Zhu '25): pass@1 and pass@32 diverge with no head-level locus to point at. We find no reorganization — but absence of evidence isn't evidence of absence.

FUTURE WORK

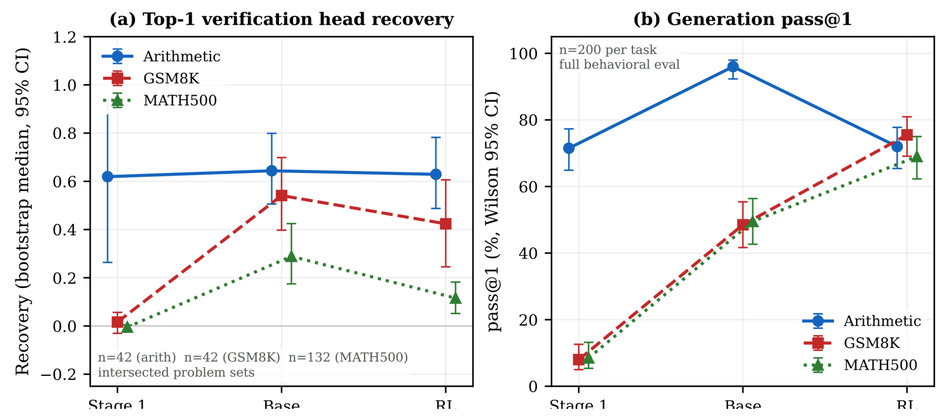
Falsifiable prediction: no softmax → no ghosts — linear-attention and state-space models should show no dissociation.
Causal test: scale the ghost head's value projection at the ghosting checkpoint.
Generality: track ghosts across other mid-training pipelines — instability general or OLMo-specific? Please, we need more open-sourced checkpoint models!
Feature view: do ghost heads write identifiable SAE/transcoder features, or is the misdirection feature-diffuse? How are they related on greedy vs distributional?

05 Answer-token patching

NEW METHOD · MULTISTEP

Single-token patching starves on multi-step problems. Fix: give the full chain of thought, corrupt **only the final answer**, patch the token that commits to *correct / incorrect*:

```
Q: What is 4729 + 1583?\nA: The answer is 6312 | 7414. The solution is [patch]
```



RL never raises verification recovery — whether generation falls or rises. Arithmetic: pass@1 drops 24pp under RL, verification flat. GSM8K / MATH500: pass@1 rises 27 / 20pp, verification still no higher. Validated on arithmetic: recovers `L29.H5` (.632) — arithmetic's *verification* head, distinct from the generation specialist `L18.H7`; a consistent-wrong-CoT control is 10× weaker. Mechanistic support for explicit process reward models (Lightman '23): outcome RL yields no self-verifier as a side effect.

TAKEAWAYS FOR PRACTITIONERS

- → **Patch with both metrics, always.** Report $r(g,p)$ alongside circuits — divergence flags instability before behavior shows it, and greedy-only analyses both invent heads (ghosts) and miss heads (inverse ghosts).
- → **Ghost heads are a signature of objective shifts** — attention-specific, recurrent, strong-but-misdirected. Dormancy criteria only catch them after the fact.
- → **Use paired problem sets across checkpoints.** Subset composition manufactured an apparent RL ghost trend on MATH500 that a paired bootstrap erased (P(Δ>0) = .40).

LIMITATIONS

Full temporal coverage in one family (OLMo); cross-architecture evidence is snapshot-level. Ghost criterion is threshold-heuristic (robust to sweeps, not noise-model-derived). Mechanism support is correlational. Verifier reranking underpowered (n=100). Answer-token patching validated on arithmetic only.

Zhang & Nanda '24 · Meng+ '22 · Yue+ '25 · Zhao+ '25 · Olsson+ '22 · Lightman+ '23 · OLMo-3 (AI2 '25)