CIRCUIT EMERGENCE DOES NOT ESTABLISH FUNCTIONAL EMERGENCE: REASONING DISTILLATION CAN RECRUIT LATENT ATTENTION-HEAD SIGNALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reasoning post-training can make an attention head causally important even when ablating the same head in the base model has little effect. This leaves open two possibilities: the head may have acquired new task-relevant function, or the post-trained network may have learned to use information that was already present. We study this distinction in Qwen2.5-Math-7B and DeepSeek-R1-Distill-Qwen-7B, whose attention heads align exactly. We scan all 784 query heads and test selected heads on disjoint holdouts. The clearest case, L26H25, has no detectable positive causal effect in the base model but a large effect after distillation. However, its base-model activation recovers 98.8% of the distilled causal effect after RMS matching. Token-position shuffling and a same-layer control head recover only 8.2% and 6.8%. Rolling back the local Q/K/V/O pathway removes only about 21% of the amplification, while transplanting the distilled local pathway into the base model does not create the distilled role. A 1.5B aligned pair shows both local-rewrite and reader-dominant cases. On a frozen 128-problem MATH subset, we detect no accuracy difference from matched-control ablation at this sample size, while restricted generation length increases by 16.0%. Newly causal circuit membership can therefore arise through recruitment of task-relevant signals already present in the base model.

1 INTRODUCTION

Reasoning post-training can change model behavior without changing architecture. Architecture-aligned base and post-trained models therefore let us compare the same physical component before and after training. Recent work has described attention heads that become important after reasoning post-training as emergent reasoning heads. New causal importance, however, does not show that new task-relevant information appeared inside the head.

We use *circuit emergence* for a component that becomes part of a causally important pathway, and *functional emergence* for new task-relevant information appearing at that component's output. These are different questions. A base-model head may already carry a useful signal that the base network largely ignores; post-training can make the same head important by changing how downstream computation uses that signal.

We test this hypothesis in Qwen2.5-Math-7B and DeepSeek-R1-Distill-Qwen-7B (Yang et al., 2024; Guo et al., 2025). The models are architecture-aligned, so every query head has an exact physical counterpart. We scan all 784 query heads in the distilled model, replicate selected heads on a disjoint holdout, test reasoning specificity against matched surface controls, and then separate source-side rewriting from downstream recruitment with bidirectional activation and parameter interventions.

L26H25 is the clearest case. Its causal effect is indistinguishable from zero in the base model and large after distillation, but the aligned base activation of the same physical head reconstructs nearly the entire distilled causal contribution. The effect disappears under token-position shuffling or replacement with a matched control head. Conversely, transplanting the distilled local Q/K/V/O pathway into the base model does not create the distilled causal role. This points to downstream

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

recruitment of a task-relevant signal that was already present in the base activation. This is not the only trajectory we observe: in the aligned 1.5B pair, some heads depend much more strongly on local Q/K/V/O rewriting.

Our contributions are:

- a longitudinal causal framework that separates circuit membership from component-level task-relevant signal content;
- evidence that reasoning-specific heads can become strongly causal after distillation even when their useful signal was already present in the base activation;
- bidirectional activation and parameter tests that distinguish local rewriting from downstream recruitment; and
- behavioral evidence that disrupting the recruited heads increases generation cost, while no final-accuracy difference is detected relative to matched controls on our 128-problem subset.

2 Related Work

Mechanisms across fine-tuning. Prakash et al. (2024) show that mathematical fine-tuning can enhance an existing entity-tracking mechanism. Chhabra et al. (2025) find recovery and amplification of task mechanisms after retraining, while Wang et al. (2025) report substantial circuit-edge reorganization despite high node overlap during mathematical fine-tuning. These works establish that fine-tuning can preserve and strengthen pre-existing mechanisms. We study the harder apparent-emergence case: a component that looks absent by causal importance before post-training.

Reasoning-model mechanisms. Zhang et al. (2025) identify Reasoning-Focus Heads in distilled DeepSeek-R1 models and establish causal influence from reasoning tokens to answers. Galichin et al. (2026) identify SAE features associated with uncertainty, exploration, and reflection. Most directly, Park et al. (2026) report attention heads that become active after reasoning post-training and describe them as novel, specialized reasoning heads. Our central point is that newly entering a causal circuit is evidence that a component is *used*, but not by itself that its task-relevant content is *new*. Cross-model transfer of the same physical component provides a direct test of these possibilities.

Circuit discovery and causal testing. Automated Circuit Discovery (ACDC) searches model graphs for compact circuits using activation-based interventions (Conmy et al., 2023), while attribution patching provides a cheaper approximation for ranking important edges (Syed et al., 2024). Causal scrubbing tests a proposed mechanistic abstraction by resampling activations that the hypothesis predicts should be interchangeable (Chan et al., 2022). These tools answer a different question from ours. ACDC aims to recover a compact whole-model circuit, and causal scrubbing tests the faithfulness of a stated circuit-level hypothesis. Here we exploit exact architectural alignment to ask how the task-relevant signal in the *same physical attention head* changes across distillation. We first rank all aligned heads, then intervene directly on matched components using cross-model activation transfer and local parameter rollback or transplant. The approaches are complementary rather than interchangeable; our head-level scan can still miss distributed mechanisms whose effect is visible only jointly.

3 Experimental Setup

3.1 Models and causal metric

We compare Qwen2.5-Math-7B with DeepSeek-R1-Distill-Qwen-7B. Both have 28 layers, 28 query heads per layer, and 4 grouped-query key/value heads. For a prompt x and continuation $y = (y_1, \ldots, y_k)$, we score the continuation by mean token log-likelihood,

$$s(y \mid x) = \frac{1}{k} \sum_{i=1}^{k} \log p(y_i \mid x, y_{<i}).$$

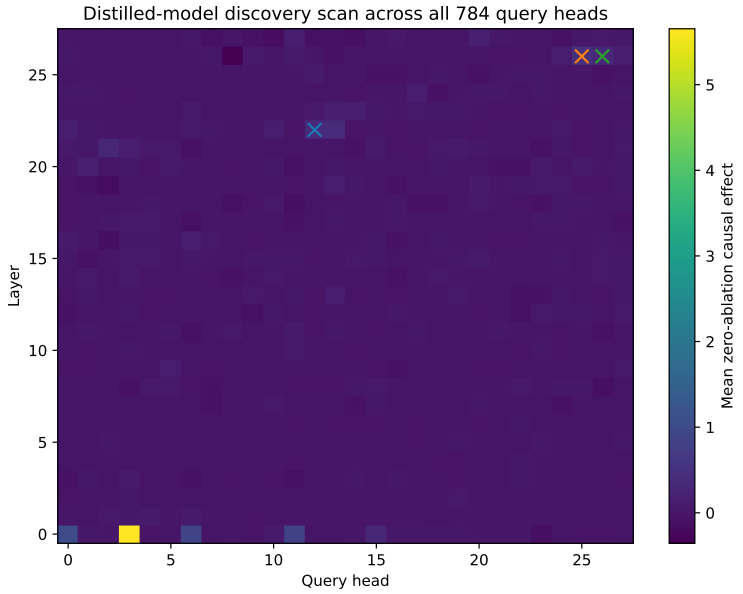


Figure 1: Full distilled-model discovery scan over all 784 query heads. Crosses mark the three reasoning-specific heads used in the primary mechanism and behavioral experiments.

Correct and matched-wrong continuations have equal token length within each probe. The correct-minus-wrong margin is

$$m = s(y^+ \mid x) - s(y^- \mid x),$$

and the causal effect of head h under ablation is

$$E_h = m_{\text{clean}} - m_{\text{ablated}}.$$

3.2 Discovery, replication, and specificity

Discovery uses an automated exhaustive single-head scan over all $28 \times 28 = 784$ query heads on 32 balanced arithmetic probes (eight each for addition, subtraction, multiplication, and division). Each probe ends at an intermediate arithmetic step and compares a correct continuation with a token-length-matched incorrect continuation; for example, a prefix ending in "$13 + 5 =$" is scored against the correct and matched-wrong number continuations. We select the top eight by distilled discovery effect only, preventing selection on the base-to-distilled difference. The 32-probe scan is a ranking screen rather than a coverage claim: narrow or context-dependent heads can be false negatives, while all selected-head claims are evaluated on separate data. The same physical heads are evaluated on a disjoint 192-probe holdout, with 12 same-layer controls and both zero and mean ablation.

To distinguish reasoning-specific effects from generic number handling, we additionally construct 96 reasoning examples and three matched 96-example controls: numeric copying, key-value lookup, and repetition of an explicitly supplied result. Correct and wrong number tokens are matched across categories. The specificity gap is reasoning causal effect minus mean control effect. Selection and evaluation are separated throughout: discovery uses only distilled-model effects, the 192-probe holdout is disjoint from discovery, and the later specificity and mechanism probe sets are frozen separately. The 64-probe mechanism set also excludes the natural reasoning examples used to construct the specificity probes.

3.3 Source rewrite versus downstream recruitment

We freeze L22H12, L26H25, and L26H26 before mechanism analysis. On 64 additional balanced probes, we patch activations bidirectionally across models. For base→distilled patching, the distilled

head activation is replaced with the same physical base-head activation while keeping the distilled downstream network. We test both raw and RMS-matched activation transfer. Two negative controls preserve important nuisance factors: token-position shuffling of the same base activation, and replacement with a pre-specified same-layer control head. The token-shuffle control is the cleaner test of example- and position-specific information because a different head can use a different feature basis.

For base→distilled activation transfer, let m_{clean}, m_{zero}, and m_{src} denote the distilled margins under the clean head, a zeroed head, and a patched base-head activation. Transfer retention is

$$R = \frac{\mathbb{E}[m_{\text{src}} - m_{\text{zero}}]}{\mathbb{E}[m_{\text{clean}} - m_{\text{zero}}]}.$$

We report this ratio for the raw base activation and after RMS matching. For RMS matching, each correct or wrong sequence is scaled independently over all non-padding token positions and head dimensions by $\alpha = \text{RMS}(a_{\text{dist}})/\text{RMS}(a_{\text{base}})$ before patching.

We also patch parameters bidirectionally. Distilled←base rollback replaces Q/K/V/O slices with base values; base←distilled transplant performs the reverse. Because the model uses GQA, Q/O interventions are query-head-specific while K/V interventions modify shared channels. Let E_B and E_D be mean base and distilled causal effects, E_{RB} the distilled effect after local rollback, and E_{TR} the base effect after the reverse transplant. We report

$$N = \frac{E_D - E_{\text{RB}}}{E_D - E_B}, \qquad S = \frac{E_{\text{TR}} - E_B}{E_D - E_B}$$

as necessity- and sufficiency-style fractions. Because transformer components interact nonlinearly, these fractions are not an additive decomposition: in particular, $1 - N$ is not interpreted as a unique downstream-reader fraction.

As a secondary diagnostic, downstream gradient sensitivity is the RMS norm of $\nabla_a m$ with respect to the selected head's pre-o_{proj} activation a, computed over the correct and wrong sequences, all non-padding token positions, and head dimensions. We report the paired distilled/base ratio on a frozen 32-probe subset.

3.4 Behavioral evaluation

We evaluate a frozen 128-problem subset drawn uniformly without replacement from MATH-500 (Hendrycks et al., 2021; Lightman et al., 2023) using a fixed random seed and greedy generation. In the primary intervention, the three reasoning heads are ablated simultaneously; the comparator simultaneously ablates three pre-specified matched controls with the same layer composition. An initial 1024-token run confounded correctness with truncation; all truncated examples were therefore rerun at 2048 and then 4096 tokens.

Because some trajectories still hit 4096, our primary efficiency metric is restricted generation length

$$\text{RGL}_{4096} = \mathbb{E}[\min(T, 4096)],$$

where T is completion length. This quantity is observed for every example without assuming that capped trajectories completed.

3.5 Statistical testing

All uncertainty calculations are paired at the probe or problem level. We use nonparametric paired bootstrap confidence intervals for mean effects and ratios, sign-flip randomization tests for paired continuous contrasts, and exact paired tests for binary correctness or threshold-crossing outcomes. Families of pre-specified related tests are corrected with Holm's procedure. The three primary 7B heads are frozen before mechanism and behavioral analysis.

4 Results

4.1 Reasoning-specific heads become newly causal

The discovery scan reveals a sparse set of high-effect heads, with strong late-layer candidates (Fig. 1). All eight selected heads replicate under zero ablation on the 192-probe holdout and six replicate

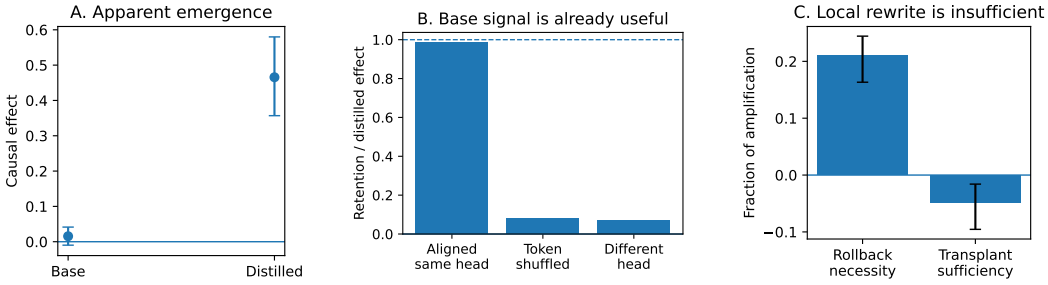


Figure 2: **L26H25 becomes newly causal while its useful signal is pre-existing.** (A) Its causal role is absent within uncertainty in the base model and large after distillation. (B) The aligned base activation reconstructs nearly all of the distilled causal contribution, unlike token-shuffled and different-head controls. (C) Rolling back the full local Q/K/V/O pathway explains only about 21% of amplification, while transplanting distilled local parameters into the base model is not sufficient to create the role.

Head	Base E	Dist. E	Spec. gap	Raw ret.	RMS ret.
L22H12	0.094	0.350	0.357	1.003	0.942
L26H25	0.016	0.466	0.344	0.984	0.989
L26H26	0.096	0.363	0.252	1.195	1.147

Table 1: Primary 7B heads. Base/distilled effects are from the 192-probe holdout; specificity is from matched surface controls; raw and RMS retention are base→distilled activation transfer on a separate 64-probe mechanism set. Retention is a signed effect ratio, so values above 1 indicate overshoot rather than a probability above 100%.

under mean ablation. Three late heads also pass the reasoning-specificity controls and are frozen for subsequent experiments.

L26H25 is the strongest apparent-emergence case. On the main holdout its base effect is 0.016 with a confidence interval crossing zero, while its distilled effect is 0.466. On the separate mechanism set, the same qualitative pattern is sharper: base effect -0.007 and distilled effect 0.480.

4.2 The base head already carries the useful signal

Figure 2 follows L26H25 through three tests. First, it becomes strongly causal after distillation. Second, replacing its distilled activation with the aligned *raw* base activation preserves 98.4% of the distilled causal contribution (95% CI 83.8–110.9%); RMS matching changes this only slightly, to 98.9% (95% CI 85.1–110.9%). Thus the transfer does not depend on artificially rescaling a weak source activation. Under RMS matching, token-position shuffling retains only 8.2%, and a same-layer matched control head retains 6.8%. Both aligned-versus-control contrasts survive Holm correction ($p = 6 \times 10^{-5}$). L22H12 shows 100.3% raw and 94.2% RMS-matched retention; L26H26 shows 119.5% raw and 114.7% RMS-matched retention.

The result does not mean the representation is unchanged. Base/distilled activation cosine is 0.53 for L26H25, and 0.70 and 0.56 for L22H12 and L26H26. Rather, the task-relevant, token-aligned signal is already accessible in the base activation even though the base network does not causally depend on it.

4.3 Local rewriting contributes but is not sufficient

For L26H25, full Q/K/V/O rollback removes 21% of causal amplification and the reverse transplant has sufficiency -4.8%. For L22H12 and L26H26, rollback removes 23% and 27%, while transplant sufficiency is 13% and 1%. Because K/V are shared within each seven-query-head GQA group, these full-pathway interventions include shared-channel changes and are not strictly single-query-head-local. Thus local parameter changes contribute to the post-training state, but the new causal role

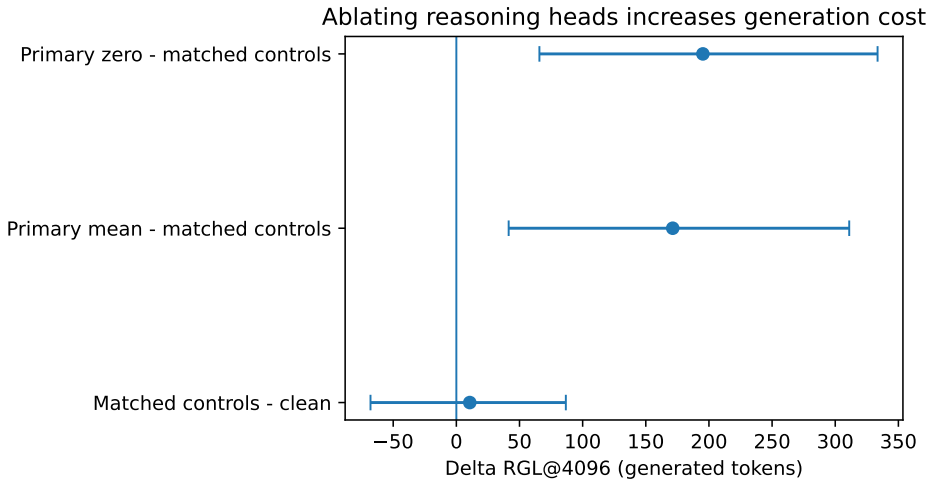


Figure 3: Primary-head ablation increases restricted generation length relative to matched controls, while matched-control ablation does not detectably differ from clean generation.

cannot be reproduced by moving the local pathway alone. We do not read the remaining 79% for L26H25 as a literal downstream share; nonlinear interactions prevent such an additive decomposition. The intervention instead establishes the narrower claim that local rewrite alone is insufficient to explain or recreate the distilled causal role.

Downstream gradient sensitivity to the primary heads increases by roughly 1.67–1.83$\times$ after distillation. Matched late-layer control heads also increase by roughly 1.65–1.75$\times$, so gradient amplification alone is not selective evidence. The stronger evidence is causal transfer: the correctly aligned base-head signal works in distilled readers, whereas scale-preserving alignment controls do not.

4.4 Ablation increases generation cost; no accuracy difference is detected

After rerunning truncated cases up to 4096 tokens, primary-head zero ablation and matched-control zero ablation both achieve 82.8% accuracy. The paired accuracy difference (matched controls minus primary heads) is 0.0 percentage points with a 95% bootstrap interval of $[-4.7, 4.7]$ points; there are five discordant examples in each direction. We therefore do not claim accuracy equivalence, only that no difference is detected at $N = 128$. The earlier apparent accuracy drop at the 1024-token cap was a truncation artifact and is not used as evidence.

However, primary-head ablation increases generation cost. Relative to matched controls, zero ablation increases RGL_{4096} by 195 tokens (95% CI 66–334), or 16.0% (95% CI 5.2–28.9%; Holm-adjusted $p = 0.008$). Mean ablation gives a consistent 171-token increase (95% CI 41–311; Holm-adjusted $p = 0.013$). Matched-control ablation differs from clean generation by only 11 tokens (95% CI -68 to 87; $p = 0.79$).

This pattern is consistent with compensation. At this sample size we do not detect a final-accuracy difference, while removing the identified heads makes the model spend more generated tokens reaching an answer.

4.5 A smaller aligned pair shows multiple trajectories

We also analyze Qwen2.5-Math-1.5B $\rightarrow$ DeepSeek-R1-Distill-Qwen-1.5B. Four candidate heads replicate on 256 external GSM8K reasoning probes (Cobbe et al., 2021). Two heads are mixed local-rewrite/reader cases: full local rollback explains roughly 65–67% of amplification while downstream sensitivity also rises. Two others are reader-dominant: local rollback explains roughly 25% and 0%, while downstream sensitivity rises by about 1.6$\times$. Thus post-training need not follow one

mechanistic trajectory; the 7B latent-recruitment result is one particularly clear member of a broader heterogeneous picture. Full 1.5B results are in Appendix B.

5 Discussion

The central result is a mismatch between causal importance and information availability. L26H25 is almost irrelevant under base-model ablation but strongly reasoning-specific after distillation. Its representation does change, yet the base activation still contains nearly all of the aligned signal that the distilled downstream network can use. The result is therefore about pre-existing usable signal, not an unchanged head.

This creates a simple failure mode for comparative circuit analysis. A component can cross an importance threshold because its readers changed, even if useful information was already present at its output. For architecture-aligned models, component importance is therefore more informative when paired with same-component activation transfer and controls that break token or head alignment without simply changing scale.

On our MATH subset, we do not detect a final-accuracy difference between primary-head and matched-control ablation. The longer trajectories under primary-head ablation are consistent with compensation. This is consistent with distillation routing existing information through a lower-generation-cost pathway, although restricted generation length cannot distinguish extra deliberation from verbosity or looping.

6 Limitations

The main latent-signal analysis covers one 7B architecture-aligned pair and three pre-specified reasoning-specific heads. The 1.5B pair replicates causal amplification and mechanism heterogeneity, but not the full 7B negative-control suite. Our mechanism probes focus on arithmetic continuations, while behavioral evaluation uses 128 randomly sampled MATH-500 problems; this sample is too small to establish a tight equivalence bound on accuracy. Cross-model activation patching is especially clean here because architectures and token positions align exactly; extending the method to non-isomorphic models is harder. Likewise, the different-head negative control is conservative rather than definitive because heads need not share a feature basis; the same-head token-shuffle control more directly tests whether the transferred effect depends on the aligned signal.

GQA limits parameter attribution because K/V channels are shared among seven query heads. We therefore label K/V interventions as shared-channel effects. Some MATH trajectories remain censored at 4096 tokens, motivating restricted generation length rather than raw completion length. RGL measures generation cost, not the semantic quality of extra tokens, and therefore does not distinguish additional deliberation from verbosity, looping, or delayed termination. Whole-head interventions also do not resolve polysemantic subfeatures; our claim concerns task-relevant information present in head outputs, not a unique monosemantic head function. We do not test whether new task-relevant transformations emerge elsewhere in the network, including MLPs; our claim is local to the information carried by the analyzed attention-head outputs. Finally, the experiments show that circuit emergence need not imply functional emergence; they do not claim that all apparent-emergence cases have this explanation.

7 Conclusion

New causal importance after distillation does not by itself identify where new functionality was learned. For L26H25, the base activation already contains a token-aligned signal that the distilled network can use, while the distilled local Q/K/V/O pathway is not sufficient to create the role in the base model. The 1.5B pair shows that other heads can follow different mixtures of local rewriting and downstream recruitment. On MATH, ablating the three primary heads raises generation length while we detect no final-accuracy difference relative to matched controls at $N = 128$. These results motivate longitudinal tests of component outputs rather than treating post-training circuit membership as direct evidence of a newly created local function.

Ethics Statement

This work analyzes publicly released language models and standard mathematical reasoning benchmarks. It involves no human subjects and no collection of personal data. The main dual-use consideration is that mechanistic understanding can inform both model auditing and model optimization; we are not aware of a more specific risk introduced by the experiments reported here.

Reproducibility Statement

We provide exact frozen probe sets, selected head identities, matched controls, raw per-example intervention outputs, bootstrap/randomization analysis code, and the staged MATH rerun procedure in the supplementary material. All candidate selection precedes the holdout, specificity, mechanism, and behavioral analyses described in the paper.

AI Use Statement

Generative AI tools were used as assistants for code drafting and debugging, statistical-analysis scripting, literature search, language editing, and feedback on experimental design, controls, and interpretation. The authors chose the research questions and experiments, made the final methodological decisions, ran the experiments, inspected the outputs, and checked the reported numbers against saved results. All AI-assisted code and text were reviewed by the authors, who take responsibility for the manuscript and results.

References

Lawrence Chan, Adrià Garriga-Alonso, Nicholas Goldowsky-Dill, Ryan Greenblatt, Jenny Nitishinskaya, Ansh Radhakrishnan, Buck Shlegeris, and Nate Thomas. Causal scrubbing: A method for rigorously testing interpretability hypotheses. AI Alignment Forum, 2022.

Vishnu Kabir Chhabra, Ding Zhu, and Mohammad Mahdi Khalili. Neuroplasticity and corruption in model mechanisms: A case study of indirect object identification. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 3099–3122, 2025. doi: 10.18653/v1/2025.findings-naacl.170.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Rei Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

A. V. Galichin, A. Dontsov, P. Druzhinina, A. Razzhigaev, O. Rogov, E. Tutubalina, and I. Oseledets. 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. doi: 10.1609/aaai.v40i36.40334.

Daya Guo, Dejian Yang, Haowei Zhang, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638, 2025. doi: 10.1038/s41586-025-09422-z.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Yein Park, Minbyul Jeong, and Jaewoo Kang. Thinking sparks!: Emergent attention heads in reasoning models during post training. *arXiv preprint arXiv:2509.25758*, 2026.

Nikhil Prakash, Tamar Rott Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. Fine-tuning enhances existing mechanisms: A case study on entity tracking. In *International Conference on Learning Representations*, 2024.

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 407–416, 2024. doi: 10.18653/v1/2024.blackboxnlp-1.25.

Xu Wang, Yan Hu, Wenyu Du, Reynold Cheng, Benyou Wang, and Difan Zou. Towards understanding fine-tuning mechanisms of LLMs via circuit analysis. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 63088–63112, 2025.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. Qwen2.5-Math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024.

Jue Zhang, Qingwei Lin, Saravan Rajmohan, and Dongmei Zhang. From reasoning to answer: Empirical, attention-based and mechanistic insights into distilled DeepSeek-R1 models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 3985–4002, 2025. doi: 10.18653/v1/2025.emnlp-main.198.

A Additional 7B results

A.1 Reasoning specificity

The full specificity analysis includes the amplified but non-primary head L22H13 and all matched surface controls. The main text retains only the three heads that show clean reasoning-specificity amplification.

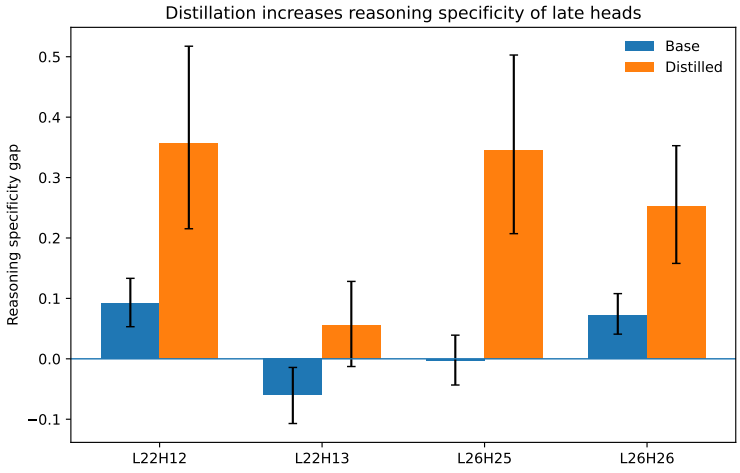


Figure 4: Reasoning specificity gap in base and distilled models.

A.2 Local Q/K/V/O interventions

Component-wise rollback and transplant results, including the GQA-shared K/V slices, are provided in the supplementary archive. The compact full-path summary is at `key_results/7b_mechanism_summary.csv`. The full component-wise table is `PARAMETER_INTERVENTION_SUMMARY.csv` inside `experiments/08_7b_mechanism/results.zip`.

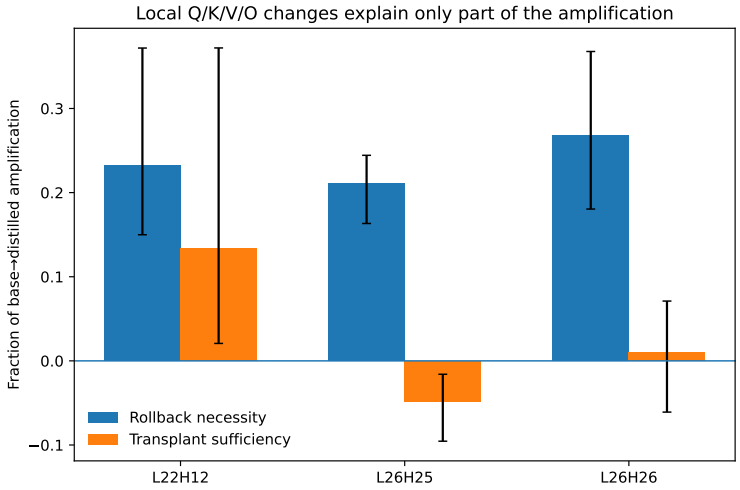


Figure 5: Full local-pathway rollback necessity and transplant sufficiency for the three primary heads.

A.3 Why the original 1024-token run was misleading

Primary-head ablation increases the probability of crossing the old 1024-token cap. After extending the generation ceiling, the apparent correctness effect disappears.

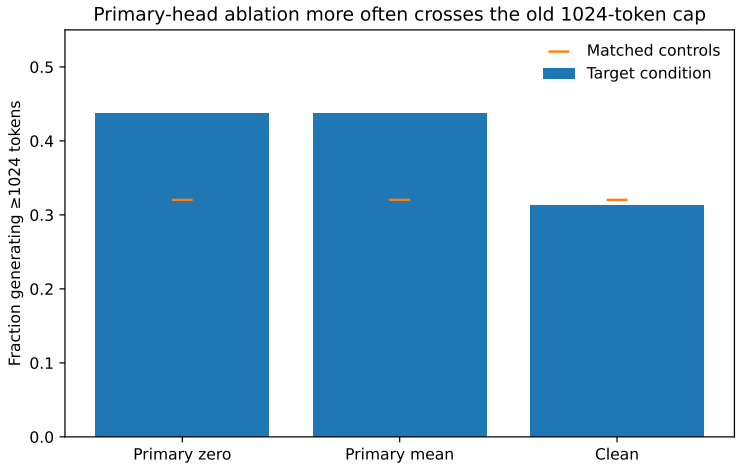


Figure 6: Fraction of examples reaching at least 1024 generated tokens. Blue bars are the labeled conditions (primary zero ablation, primary mean ablation, or clean generation); the orange dash is the same matched-control zero-ablation reference repeated in each column. The clean bar contains no head ablation.

B 1.5B Replication

On the 1.5B aligned pair, L26H10, L26H6, L22H6, and L22H8 all show significant causal amplification on 256 external GSM8K probes. Mechanistic follow-up separates mixed local-rewrite/reader cases from reader-dominant cases.

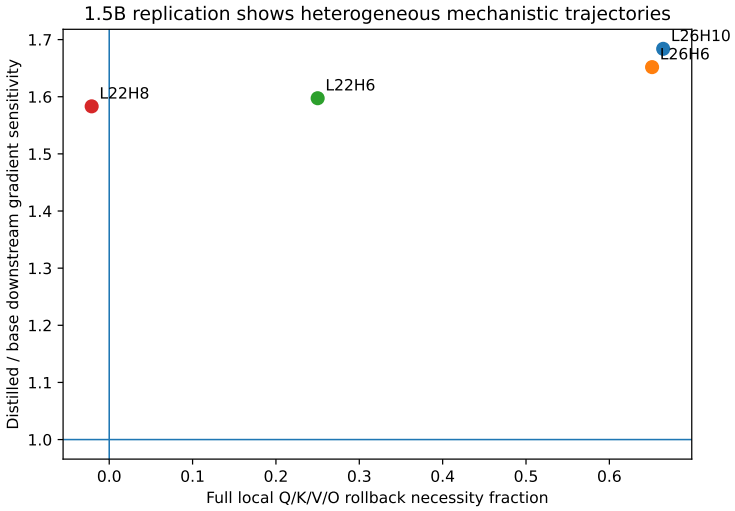


Figure 7: 1.5B mechanism heterogeneity: local rollback necessity versus downstream sensitivity amplification.