Mechanistic origins of catastrophic forgetting: why RL preserves circuits better than SFT?

Jeanmely Rojas Nunez [1] **Viraj Sawant** [1] **Nathan Allen** [1] **Nomgondalai Amgalanbaatar** [1] **Yannis Zongo** [1]
Vasu Sharma [1] **Maheep Chaudhary** [2]

Abstract

Fine-tuning large language models (LLMs) frequently induces catastrophic forgetting of prior capabilities. Recent work has shown that reinforcement learning (RL) can retain prior capabilities more effectively than supervised fine-tuning (SFT), and this effect is often interpreted through policy proximity to the base policy. We revisit this behavioral account at the mechanistic level and ask whether RL's advantage is associated with stronger preservation of internal computational circuits, even when output-space KL alone does not fully predict forgetting. We introduce differential circuit vulnerability, a head-level measure of how much a circuit degrades under fine-tuning, and use it to compare RL and SFT on Qwen2.5-3B-Instruct adapted to scientific question-answering. We find a mechanistic trade-off in this setting: SFT adapts more rapidly to the target task but produces greater circuit disruption and forgetting of prior capabilities, whereas RL preserves a larger fraction of the base circuit at the cost of slower task adaptation. These findings suggest that circuit preservation may help explain RL's robustness to catastrophic forgetting in our experiments. Our code is available at: https://github.com/rl-sft-circuit-research/differential-circuit-vulnerability.

1. Introduction

Adapting large language models (LLMs) to new downstream tasks often leads to catastrophic forgetting: gains on a new objective come at the expense of prior capabilities. As models are increasingly expected to update continuously and adapt to new domains, mitigating such degradation has become a central challenge in post-training. Prior work establishes that the choice of adaptation objective strongly shapes this trade-off: RL often preserves prior capabilities more effectively than SFT (Shenfeld et al., 2025; Hu et al., 2025). This advantage is commonly interpreted through policy proximity to the pretrained model, often measured by KL divergence. However, output-space proximity may not fully explain retention in all settings. We therefore revisit this account mechanistically and ask whether retention is better aligned with preservation of internal circuits than with output-space KL alone.

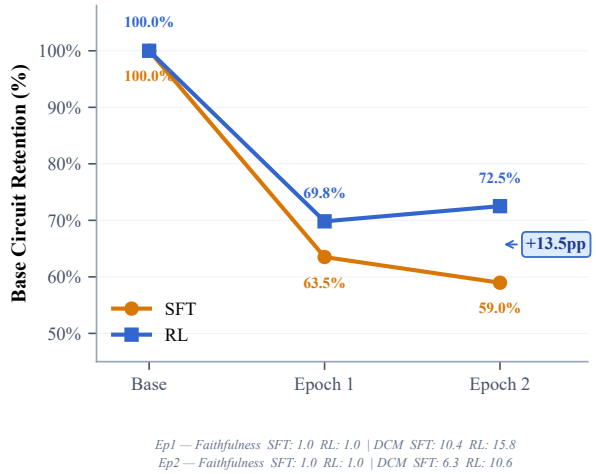

Figure 1. **Circuit retention trajectories during high-NTS training.** Starting from 100% base-circuit retention, SFT (orange) and RL (blue) diverge over two training epochs. SFT drops to 63.5% after Epoch 1 and further to 59.0% by Epoch 2. RL drops to 69.8% after Epoch 1 and increases to 72.5% by Epoch 2, a 13.5 percentage-point difference at Epoch 2. Footer values report per-epoch circuit faithfulness, both $\approx$ 1.0, and Differential Causal Mediation (DCM). RL yields higher DCM at both epochs: 15.8 vs. 10.4 at Epoch 1 and 10.6 vs. 6.3 at Epoch 2, indicating that the retained circuit components exert greater causal influence on model outputs under RL than under SFT.

In parallel, mechanistic interpretability has shown that model capabilities are implemented by internal computational *circuits* composed of attention heads, MLP layers,

[*]Equal contribution [1]Algoverse AI Research [2]Independent. Correspondence to: Jeanmely Rojas Nunez <jeanmelyrojas-nunez@gmail.com>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

and residual pathways. Fine-tuning may therefore succeed or fail depending on whether it preserves these structures or disrupts them (Prakash et al., 2024). This motivates our central question: **is RL's retention advantage associated with stronger preservation of task-relevant circuits relative to SFT?**

We examine post-training through the lens of circuit preservation. We introduce *differential circuit vulnerability*, the relative susceptibility of internal computational subgraphs to degradation under different training objectives. Using this framework, we compare RL and SFT on matched adaptation tasks and identify which circuits are preserved or disrupted. Figure 1 previews the central finding in our setting: over the course of training the two objectives trace diverging circuit-retention trajectories, with RL preserving a larger fraction of the base circuit. We characterize this trade-off mechanistically in the sections that follow.

We instantiate our analysis on Qwen2.5-3B-Instruct, independently adapting the base model via SFT and RL on scientific question answering and evaluating retention across a broad benchmark suite spanning commonsense reasoning, factuality, and instruction following. The results expose a trade-off in this setting: SFT adapts faster to the new objective but at the cost of greater circuit disruption, whereas RL preserves more of the base circuit and of prior capabilities, occasionally at the expense of under-optimizing the new task.

These results suggest that continual adaptation may be usefully understood not merely as parameter optimization, but as the *selective preservation and modification* of circuits encoding skills, offering a possible mechanistic account of RL's greater conservatism during post-training.

2. Related Work

Reinforcement Learning, Supervised Fine-Tuning, and Forgetting. Supervised fine-tuning (SFT) and reinforcement learning (RL) are the dominant post-training methods for adapting foundation models, yet their effects on previously acquired capabilities are not fully understood. SFT optimizes imitation on a target distribution, whereas RL updates behavior through reward optimization and often produces qualitatively different parameter changes. Recent studies suggest that, under comparable settings, RL can preserve pretrained capabilities better than SFT, these findings indicate that the training objective influences both adaptation and forgetting (Shenfeld et al., 2025; Hu et al., 2025).

Dr.GRPO. The RL method used in this work, Dr.GRPO, builds on Group Relative Policy Optimization (Zhihong et al., 2024), a policy gradient algorithm that estimates advantages by normalizing rewards within a sampled group of completions rather than relying on a learned value function.

Mechanistic Interpretability and Circuit Analysis. Mechanistic interpretability explains model behavior through concrete components such as attention heads, MLP layers, and circuits. Intervention-based methods, including activation patching, path patching, and masking, have shown that many behaviors can be localized to sparse causal subnetworks, and recent work provides a unifying causal-abstraction framework for these techniques (Geiger et al., 2025). Related work further suggests that fine-tuning frequently modifies existing mechanisms rather than replacing them entirely (Davies et al., 2023; Prakash et al., 2024), that training procedures themselves can be shaped to yield more modular circuit structure (Golechha et al., 2025), and that some circuits are alignment-critical and should be preferentially preserved during model modification (Patel et al., 2025).

Evaluating Retention Across Capabilities. Retention cannot be assessed on the adaptation task alone. Models may improve on a new objective while degrading reasoning, factuality, instruction following, or code generation. For this reason, prior work increasingly relies on diverse benchmark suites. Common evaluations include MMLU, HellaSwag, and WinoGrande for reasoning and commonsense inference, and TruthfulQA, IFEval, and HumanEval for factuality, instruction following, and coding (Hendrycks et al., 2021; Zellers et al., 2019; Lin et al., 2022).

Research Gap and Positioning. Despite progress in both post-training and interpretability, the two literatures remain loosely connected. Optimization studies typically report benchmark outcomes without explaining the internal causes of forgetting, while interpretability studies rarely compare learning objectives. We address this gap by comparing RL and SFT through the lens of circuit preservation, asking which objective better maintains the causal structures underlying pretrained capabilities.

3. Methodology

We test whether RL's retention advantage is associated with stronger preservation of task-relevant internal circuits than SFT. Our pipeline comprises three phases: (I) compare retention, output-space drift, and circuit preservation between SFT and RL; (II) identify circuits in each model; and (III) compare how post-training reshapes these circuits.

We train two models independently from the same pretrained checkpoint π_{base}: We first train an SFT model π_{SFT} via completion-only supervision, and an RL model π_{RL} via Dr.GRPO applied directly to the base model. This parallel design provides a controlled comparison of adaptation method at matched training compute, isolating the effect of the learning signal (supervised imitation vs. reward optimization) rather than conflating it with initialization: $\pi_\theta \in \{\pi_{\text{SFT}}, \pi_{\text{RL}}\}$.

To quantify behavioral drift from the base model, we compute the expected KL divergence on retention tasks, where lower values correspond to smaller distributional shift.

$$\mathbb{E}_{x\sim\tau}\left[D_{\mathrm{KL}}\left(\pi_{\mathrm{base}}(\cdot|x)\,\|\,\pi_\theta(\cdot|x)\right)\right]. \tag{1}$$

3.1. Phase I: Reproducing Distributional Shift Effects

We first compare retention, output-space KL, and circuit preservation across SFT and RL in order to test whether behavioral retention is better aligned with policy proximity or with internal circuit preservation (Shenfeld et al., 2025). Models are trained on a downstream Task A and evaluated on a separate suite of retention benchmarks (Task B).

SFT. We fine-tune π_{base} with a completion-only cross-entropy loss.

RL. We apply Dr.GRPO directly to π_{base}. The model samples candidate completions, receives binary domain-specific rewards, computes normalized group-relative advantages, and updates the policy through the advantage-weighted log-probability objective. We use a group size of $G = 64$, two refinement steps ($\mu = 2$), and no explicit KL penalty.

3.2. Phase II: Circuit Identification via Differential Binary Masking

We analyze circuits at the attention-head level using Differential Binary Masking (DBM) (Chaudhary & Geiger, 2024). DBM learns a mask over heads that interpolates between base and counterfactual activations: $\tilde{a}_h = (1 - m_h)a_h^{\mathrm{base}} + m_h a_h^{\mathrm{source}}$

where $m_h \in [0, 1]$. Annealing pushes the masks toward binary selections, yielding sparse causal circuits.

Triplets. For chemistry QA, we construct triplets $(x_{\mathrm{base}}, x_{\mathrm{source}}, y_{\mathrm{target}})$ corresponding to three counterfactual hypotheses: answer-key swaps, molecule swaps, and task-type swaps.

Objective. Masks are optimized to increase the probability of the target answer while remaining sparse: $\mathcal{L}_{\mathrm{DBM}} = -\log P(y_{\mathrm{target}}|x, \tilde{a}) + \lambda \sum_h m_h$

Scoring. We score answers using the geometric mean of token probabilities

$$p(y|x) = \exp\left(\frac{1}{T}\sum_{i=1}^{T}\log P(y_i|x, y_{<i})\right). \tag{2}$$

Circuit discovery is run independently for π_{base}, π_{SFT}, and π_{RL}.

3.3. Phase III: Cross-Model Circuit Comparison

We assess how strongly the discovered circuits remain functional after post-training. Circuit faithfulness is defined as:

$$\mathrm{Faithfulness}(\mathcal{C}, M) = \frac{F(\mathcal{C}|M)}{F(M)}. \tag{3}$$

with values close to 1 indicating that the circuit recovers most of the model's behavior.

For each head h, we compare DBM mask values against those of the base model:

$$\Delta m_h(M) = m_h^M - m_h^{\mathrm{base}}, \tag{4}$$

where $M \in \{\pi_{\mathrm{SFT}}, \pi_{\mathrm{RL}}\}$. This identifies heads that are preserved, amplified, or weakened by training.

We define *vulnerable* heads as those more degraded under SFT than under RL: $\mathcal{C}_{\mathrm{vuln}} = \{h : m_h^{\mathrm{SFT}} < m_h^{\mathrm{RL}} - \delta\}$.

4. Experiments

Our experiments evaluate whether RL's retention advantage is associated with stronger preservation of task-relevant circuits than SFT. We address two questions: (I) How do forgetting and behavioral drift compare between SFT and RL; and (II) are retention differences reflected in stronger circuit faithfulness, more stable head-level contributions, and more distributed circuits?

4.1. Experimental Setup

All experiments use Qwen2.5-3B-Instruct. The fine-tuning task (Task A) is scientific question answering. Retention (Task B) is measured on benchmarks spanning commonsense reasoning, factuality, instruction following, and code generation, since forgetting can be capability-specific. We compare three systems: the pretrained base model, standard SFT, and RL with Dr.GRPO. Both adapted models are initialized independently from π_{base}, ensuring that the comparison isolates the learning signal rather than initialization effects. Behavior is measured with downstream accuracy and KL divergence from the base model. The mechanistic analysis uses circuit faithfulness, head-level mask shifts $\Delta m_h = m_h^M - m_h^{\mathrm{base}}$ derived from cross-model DBM mask comparison, and necessity/sufficiency interventions; necessity measures the log-probability drop when a head is disrupted; sufficiency measures how much behavior is recovered when only that head is kept. Together, these metrics distinguish distributed contributors from critical bottlenecks.

New Task Score (NTS). We define the *New Task Score* (NTS) as the model's accuracy on Task A (SciKnowEval science QA), computed as the fraction of questions for which the model assigns the highest geometric-mean token probability (Eq. 2) to the correct answer choice. Table 1 summarizes the full setup.

4.2. Results

DBM identifies a base circuit comprising 297 attention heads (51.6% of all attention heads). After adaptation, the SFT model exhibits structural compression to roughly 265 heads (46.0%), whereas the RL model retains approximately 296 heads (51.4%), close to the base model's 297. The same disparity is reflected in base-circuit overlap: the RL model preserves about 68% of base heads, substantially more than the SFT model's 52% (Figure 4).

Figure 1 traces this divergence over the course of the training runs that yield the high new-task score models. Both objectives begin from full base-circuit retention, but their trajectories separate immediately: after the first epoch SFT retains only 63.5% of base heads compared to RL's 69.8%, and by the second epoch the gap widens to 13.5 percentage points (59.0% vs. 72.5%). Notably, RL's retention *increases* between epochs, suggesting that continued RL training can recover previously disrupted circuit components rather than monotonically degrading them. Throughout training, RL also maintains consistently higher DCM scores (15.8 vs. 10.4 at Epoch 1; 10.6 vs. 6.3 at Epoch 2), indicating that RL's preserved circuit carries substantially more causal signal on the answer-key counterfactual. The two objectives thus diverge along different axes of the same trade-off: SFT collapses adaptation into a small specialist set that discards much of the base circuit, while RL spreads adaptation across a broader, more causally engaged subgraph that retains substantially more of the original computation.

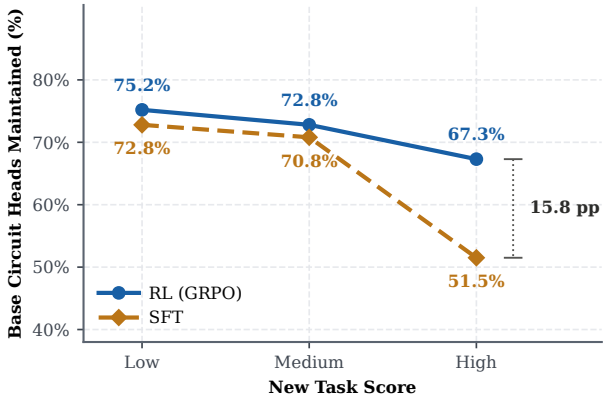

Figure 2. **Performance–preservation trade-off across NT levels.** SFT (dashed) shows a steeper decline in circuit retention as new-task score increases, while RL (solid) declines more gradually. At peak new-task performance, RL retains 15.8 percentage points more of the base circuit than SFT. Low (<50%), Medium (50–65%), High (>65%).

Circuit faithfulness exceeds 1.0 for all three models (1.02 base, 1.04 SFT, 1.12 RL), confirming that the extracted subgraphs recover the full model's behavior on the target task.

4.3. Functional Importance and Circuit Drift

Beyond aggregate circuit size, we ask whether fine-tuning preferentially modifies the heads that matter most to the base model. Figure 3 shows the necessity–sufficiency landscape: SFT produces a tight cluster of "critical specialists"—heads with both high necessity and high sufficiency—while RL maintains a flatter distribution that closely tracks the base model. This is the mechanistic signature of compression: SFT consolidates new-task behavior into a small set of high-importance heads, whereas RL spreads adaptation across many components without any single one becoming indispensable.

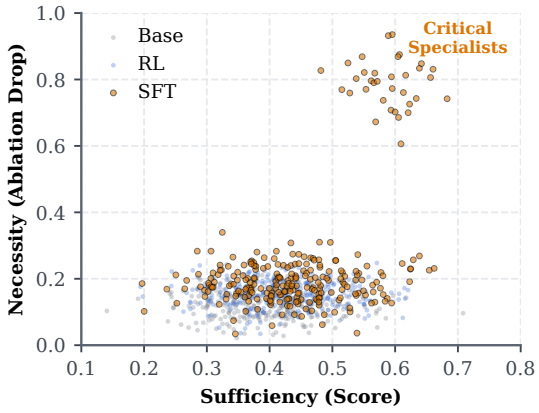

Figure 3. **Head Role Distribution Under Base, SFT, and RL Training.** SFT produces a cluster of heads with high necessity and high sufficiency scores, corresponding to the upper-right region, relative to the base model. RL yields a distribution that more closely overlaps with the base model, with fewer heads concentrated in the high-necessity, high-sufficiency region.

Figure 7 probes this directly, plotting per-head necessity against the mask shift Δm_h (Eq. 4). If either objective preferentially edited the most important heads, we would expect a positive correlation. We do not observe one: SFT shows a weak negative trend (r=−0.125, p=0.12) and RL is essentially flat (r=0.022). Read together with Figure 3, the picture is consistent across both models: neither objective rewires the circuit by tracking base-model necessity. SFT instead reorganizes broadly while elevating a new set of specialists, whereas RL adjusts uniformly across heads without producing any such concentration.

4.4. Discussion

Taken together, our findings suggest that SFT exhibits a mechanistic "breaking point" in this setting. SFT and RL are comparable at low levels of adaptation, suggesting that both objectives can improve new-task performance without substantially disrupting the base circuit. However, in the high-NTS regime, SFT shifts from conservative adaptation to aggressive circuit reconfiguration, while RL remains comparatively stable. This pattern is consistent with RL

preserving and reusing existing representations more effectively than SFT, which appears to overwrite portions of the base circuit under stronger new-task pressure. The epoch-level trajectory in Figure 1 reinforces this interpretation: SFT's retention declines monotonically across epochs, while RL retention recovers after an initial drop, suggesting that RL may consolidate and restore circuit structure during continued training. Confirming this interpretation will require experiments across additional models, tasks, and optimization settings.

We also observe that the RL model maintains greater prior-task retention and circuit preservation despite having a larger output-space Kullback–Leibler (KL) divergence than the SFT model. This suggests that output-space metrics may not reliably predict internal forgetting in this setting, this finding is consistent with broader observations that surface behavior and internal representational transfer can diverge during post-training procedures such as distillation (König et al.). In particular, RL may alter the model's output distribution while preserving much of the underlying computation responsible for prior-task performance. By contrast, SFT can appear less divergent at the output level while still reorganizing the internal circuits more aggressively.

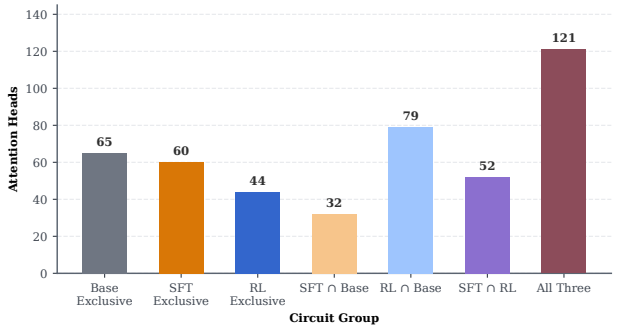

Figure 4. **Attention Head Overlap Between Base, SFT, and RL.** Each bar segment reports the number of attention heads belonging exclusively to one model, shared by exactly two models, or present in all three. RL shares more heads with the base model than SFT does, while SFT retains the fewest heads in common with either the base model or RL.

Of the 297 base-model heads, the RL model shares 200 with the base circuit compared to only 153 for SFT, a difference of 47 heads as depicted in Figure 4. SFT also has the smallest exclusive set shared with RL (52 heads), while 121 heads appear in all three models. This supports the characterization of SFT as a "compressor" and RL as a "distributed adaptor", SFT concentrates new-task behavior into a smaller, more specialized circuit, while RL preserves a broader set of base-model heads during adaptation.

5. Conclusion

We asked whether RL's retention advantage over SFT is associated with stronger preservation of internal circuits. In our Qwen2.5-3B-Instruct science-QA setting, our behavioral and mechanistic analyses are consistent with this hypothesis: RL preserves more of the original circuit and maintains stronger functional faithfulness, even when output-space KL does not indicate closer proximity to the pretrained model. By contrast, SFT adapts more quickly to the new objective but induces greater circuit reorganization and more forgetting.

These results point to a potential trade-off in post-training: rapid specialization versus stable reuse of existing mechanisms. In our setup, SFT optimizes the new objective efficiently but can disrupt circuits supporting prior skills, while RL appears more conservative, preserving those circuits at the cost of smaller task gains. More broadly, future adaptation methods may benefit from combining efficient learning with selective circuit preservation toward models that continue improving without forgetting.

6. Limitations and Future Work

This study examines only one model (Qwen2.5-3B-Instruct), which limits the generalizability of our findings. Future work should validate these trade-offs across diverse architectures (Gemma, Mistral, Llama, and Pythia) and at varying parameter scales. Our circuit analysis is further constrained to attention heads and a narrow task set. Expanding to MLP layers, residual-stream features and broader capability domains – multilingual reasoning, factual recall, safety, and tool use – would yield a more complete mechanistic picture.

References

Chaudhary, M. and Geiger, A. Evaluating open-source sparse autoencoders on disentangling factual knowledge in gpt-2 small. *arXiv preprint arXiv:2409.04478*, 2024.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. d. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., Ryder, N., Pavlov, M., Power, A., Kaiser, L., Bavarian, M., Winter, C., Tillet, P., Such, F. P., Cummings, D., Plappert, M., Chantzis, F., Barnes, E., Herbert-Voss, A., Guss, W. H., Nichol, A., Paino, A., Tezak, N., Tang, J., Babuschkin, I., Balaji, S., Jain, S., Saunders, W., Hesse, C., Carr, A. N., Leike, J., Achiam, J., Misra, V., Morikawa, E., Radford, A., Knight, M., Brundage, M., Murati, M., Mayer, K., Welinder, P., McGrew, B., Amodei, D., McCandlish, S., Sutskever, I., and Zaremba, W. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Davies, X., Nadeau, M., Prakash, N., Shaham, T. R., and

Bau, D. Discovering variable binding circuitry with desiderata. *arXiv preprint arXiv:2310.02336*, 2023.

Feng, K., Shen, X., Wang, W., Zhuang, X., Tang, Y., Zhang, Q., and Ding, K. Sciknoweval: Evaluating multi-level scientific knowledge of large language models. *arXiv preprint*, 2025.

Geiger, A., Ibeling, D., Zur, A., Chaudhary, M., Chauhan, S., Huang, J., Arora, A., Wu, Z., Goodman, N., Potts, C., et al. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025.

Golechha, S., Chaudhary, M., Velja, J., Abate, A., and Schoots, N. Modular training of neural networks aids interpretability. *arXiv e-prints*, pp. arXiv–2502, 2025.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2021.

Hu, J., Zhang, Y., Han, Q., Jiang, D., Zhang, X., and Shum, H.-Y. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. *arXiv preprint*, 2025.

König, U., Kazmi, H., Li, R., and Chaudhary, M. Quantifying subliminal behavioral transfer ratios in language model distillation. In *Continual Adaptation at Scale: Towards Sustainable AI Workshop*.

Lin, S., Hilton, J., and Evans, O. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2022.

Patel, D., Gervacio, G., Raimi, D., Zhu, K., Lagasse, R., Grand, G., Panda, A., and Chaudhary, M. Alignment-constrained dynamic pruning for llms: Identifying and preserving alignment-critical circuits. *arXiv preprint arXiv:2511.07482*, 2025.

Prakash, N., Shaham, T. R., Haklay, T., Belinkov, Y., and Bau, D. Fine-tuning enhances existing mechanisms: A case study on entity tracking. In *International Conference on Learning Representations (ICLR)*, 2024.

Sakaguchi, K., Le Bras, R., Bhagavatula, C., and Choi, Y. Winogrande: An adversarial winograd schema challenge at scale. In *AAAI Conference on Artificial Intelligence*, 2020.

Shenfeld, I., Pari, J., and Agrawal, P. Rl's razor: Why online reinforcement learning forgets less. *arXiv preprint arXiv:2509.04259*, 2025.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi, Y. Hellaswag: Can a machine really finish your sentence? In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

Zhihong, S., Peiyi, W., Qiha, Z., Runxin, X., Junxiao, S., Xiao, B., Haowei, Z., Mingchuan, Z., Li, Y., Wu, Y., and Daya, G. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Zhou, J., Lu, T., Mishra, S., Brahma, S., Basu, S., Luan, Y., Zhou, D., and Hou, L. Instruction-following evaluation for large language models. In *arXiv preprint arXiv:2311.07911*, 2023.

A. Appendix

Table 1. Experimental Setup

Component	Details
Model	Qwen2.5-3B-Instruct
Task A (Fine-tuning)	SciKnowEval (Feng et al., 2025) (Science Q&A)
Task B (Retention)	HellaSwag (Zellers et al., 2019), TruthfulQA (Lin et al., 2022), MMLU (Hendrycks et al., 2021), IFEval (Zhou et al., 2023), WinoGrande (Sakaguchi et al., 2020), HumanEval (Chen et al., 2021)
Metrics	KL divergence (Eq. 1) Adaptation performance and retention accuracy Circuit faithfulness (Eq. 3) Head-level mask shift $\Delta m_h(M)$ from DBM mask comparison (Eq. 4)
Baselines	Base model (no fine-tuning) SFT (standard) RL (Dr.GRPO)

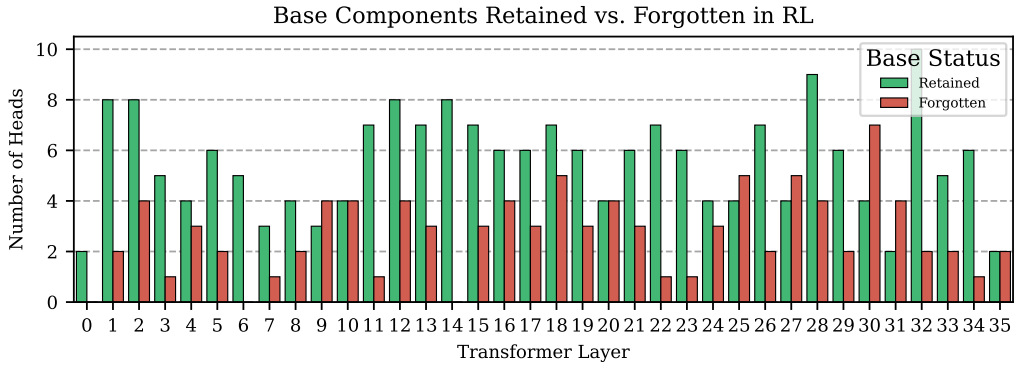

Figure 5. **Layer-wise Circuit Retention Under RL Training.** Each bar reports the number of attention heads per layer categorized as retained or forgotten relative to the base model circuit, across all 36 transformer layers. Retained heads outnumber forgotten heads in the majority of layers.

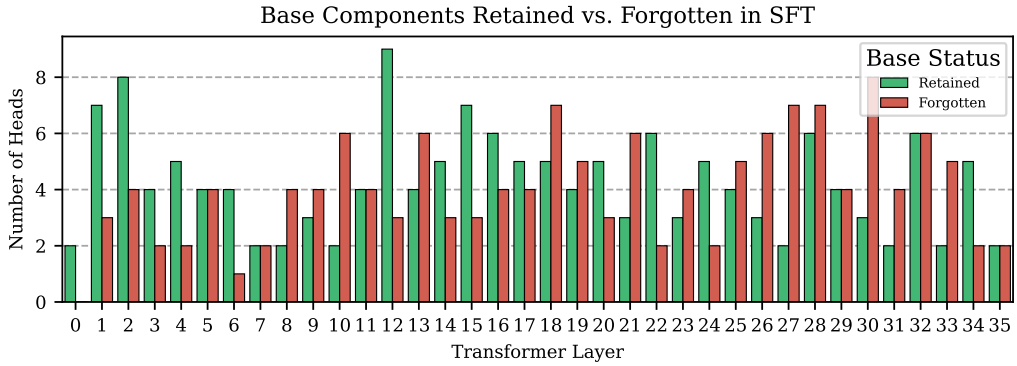

Figure 6. **Layer-wise Circuit Retention Under SFT Training.** Each bar reports the number of attention heads per layer categorized as retained or forgotten relative to the base model circuit, across all 36 transformer layers. Forgotten heads appear across all layers, with higher counts in the mid-to-late layers.

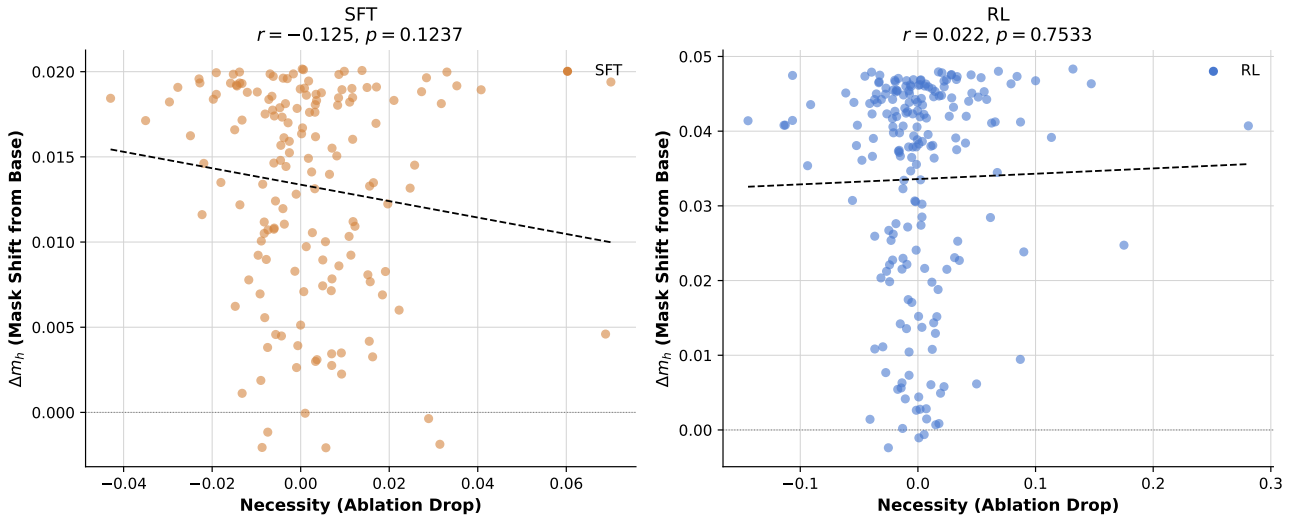

Figure 7. **Per-Head Necessity vs. Δm_h: SFT vs. RL.** Each point represents one attention head, plotting necessity score against Δm_h as defined in Eq. 4. Neither model shows a positive correlation between necessity and mask shift. SFT exhibits a weak negative correlation, $r = -0.125$, and RL shows a near-zero correlation, $r = 0.022$.

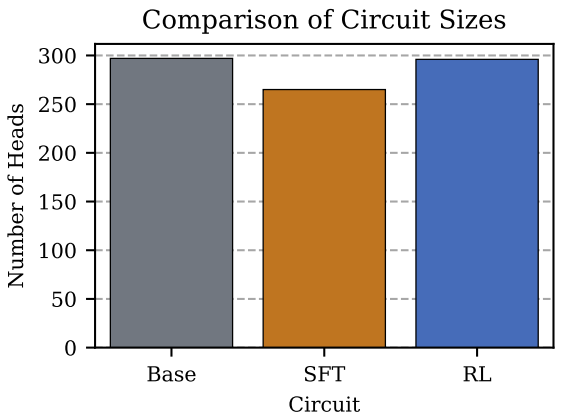

Figure 8. **Circuit Size Comparison Across Base, SFT, and RL Models.** The base model circuit comprises 297 attention heads. After adaptation, SFT compresses to approximately 265 heads (46.0% of all heads), while RL retains approximately 296 heads (51.4%), nearly identical to the base.

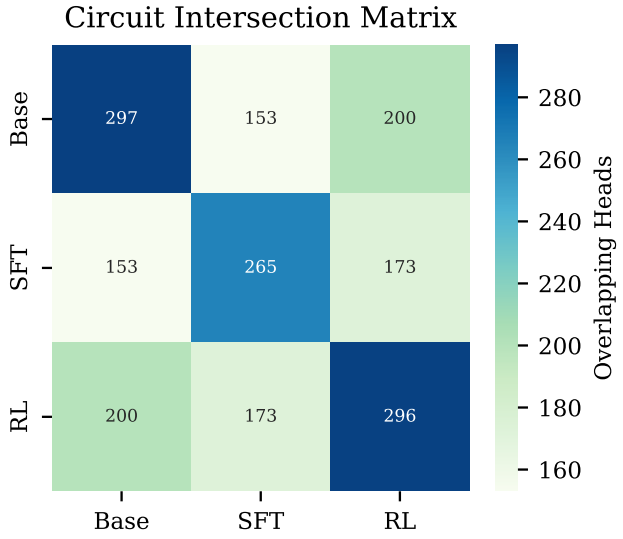

Figure 9. **Pairwise Attention Head Overlap Across Base, SFT, and RL Circuits.** Each cell reports the number of attention heads shared between two circuits. Diagonal entries report the total circuit size for each model (Base, SFT, RL). Off-diagonal entries report the number of heads common to the corresponding pair of circuits.