PREDICTIVE MECHANISTIC INTERPRETABILITY OF POST-TRAINING: A PRIORI FORECASTS OF CIRCUIT LOCALIZATION AND SIGN FROM THE BASE MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Which attention heads a post-training update will change, and in which direction, can be forecast before training, from the base model alone. In the settings we study, only how much they change requires the training run. Mechanistic accounts of post-training have so far been retrospective: circuits are read off the model after fine-tuning, predictions are scalar quantities in output space, or reachable outputs are characterized without the direction of change. We instead seal a per-head, signed forecast from the base model, fine-tune, and confirm it. **Which** heads change, and the **sign** of each change, follow under measurable base-tangent conditions and an empirically tested pathwise sign-stability condition, without requiring the update to remain small. **How much** each head changes rests on a linearization, whose validity we quantify with a creation index (the deviation of the realized update from the base-tangent prediction). These magnitude forecasts are noisier. We pre-register the which-and-sign predictions on GPT-2 small (IOI, SFT) and confirm them (localization rank correlation 0.65, per-head sign agreement 1.00), with a held-out duality check at 0.88 and causal head ablations. The predictions generalize across ten models (124M–1.5B), three circuits, and three objectives (SFT, DPO, RLVR), against composition-matched chance. The sign forecast also outlives its motivating linearization, which we call the structural–metric dissociation: it survives a DPO update with creation index 0.87 and flips as predicted (0.90 per seed) under a sealed reversal fine-tune, while magnitude calibration degrades as the creation index grows. Recent output-space laws of post-training admit the same base-model reading.

1 INTRODUCTION

A recent consensus holds that post-training largely elicits capability already latent in the base model rather than creating it: base models match or exceed RL models at large pass@k (Yue et al., 2025), RLVR is largely confined to the base support (Wu et al., 2025), and pure base-model sampling recovers RL-level reasoning (Karan & Du, 2026). All are output-space statements. In parallel, mechanistic interpretability has shown post hoc that RL edits a small (5–30%) sparse subnetwork (Mukherjee et al., 2025) and reuses latent directions already present in the base (Ward et al., 2025), but the authors of the localization result explicitly leave open where in the model that subnetwork sits.

In all of these lines the change is read off after training, the predicted quantity is a scalar in output space, or support is characterized without direction. We ask whether the circuit-level changes a post-training update makes, and the sign of each change, can be forecast before training from the base model alone (Figure 1). Our parameter-space account takes linearization as its working assumption (the lazy or NTK regime, Jacot et al., 2018; Chizat et al., 2019): during post-training, $f(x;\theta_0 + \Delta\theta) \approx f(x;\theta_0) + J_0(x)\Delta\theta$. Several predictions then follow from quantities computed on the base model before training. We make three a priori predictions (localization, reachability, forgetting), prove a signed duality between the last two, and measure the failure of the assumption as the part of the realized update the base tangent does not explain. Gradient-magnitude scores are confounded by activation scale and depth and do not localize onto interpretable circuits. Direct logit attribution (DLA) does, and it alone carries the sign of each change.

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.

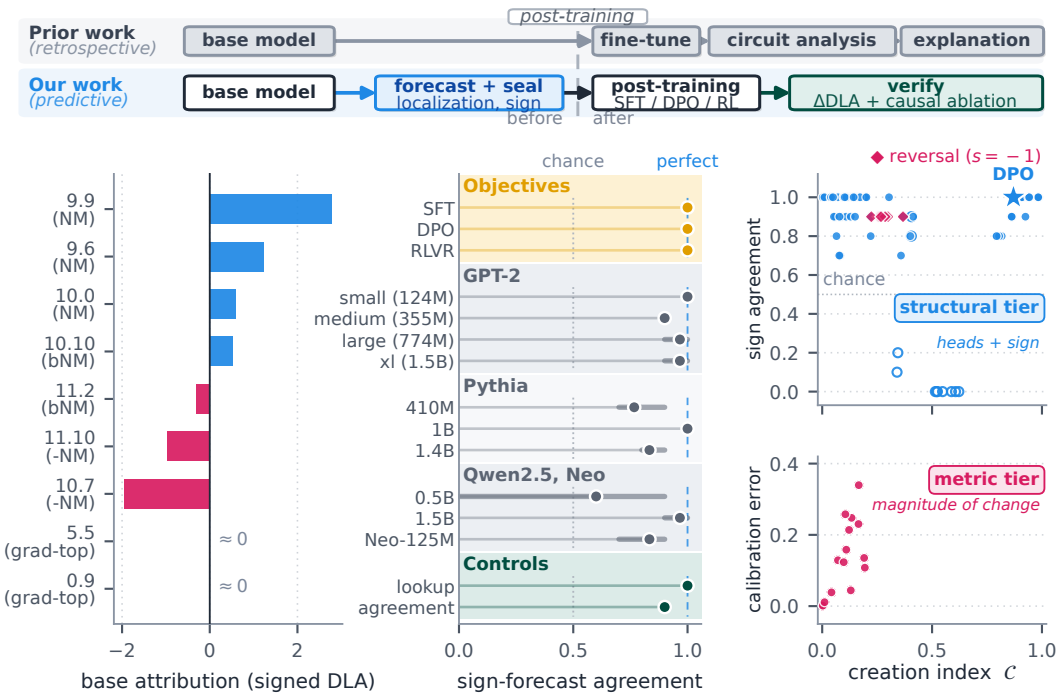

Figure 1: **Base-model forecast of circuit change and sign.** *Left:* GPT-2 IOI base attribution (NM = Name Mover, bNM = backup, −NM = negative, grad-top = naive-gradient top head). *Middle:* sign agreement per objective, model, and control (bands: groups; whiskers: seed range). *Right:* sign agreement stays high except in over-driven runs (open markers), and calibration error rises with $\mathcal{C}$.

We build on standard machinery, NTK and eNTK linearization and the Fisher-gauged reachability cost (Jacot et al., 2018; Malladi et al., 2023; Ortiz-Jimenez et al., 2023), and use it for sealed, signed, per-head forecasts made before training and confirmed after it. In the settings we study, post-training has a structural tier forecastable from the base model and a metric tier calibrated only while the update remains tangent-reachable.

Our main contributions are as follows.

- We forecast, before training and from the base model alone, which circuit components a post-training update changes and the sign of each change, and confirm these sealed predictions on real models with causal validation.
- We derive the sign forecast from the base tangent under measurable conditions (Corollary 1) and show that it outlives the linearization, holding for a DPO update with creation index $\mathcal{C} = 0.87$ and flipping under a sealed reversal fine-tune (0.90 per seed) while magnitude calibration degrades with $\mathcal{C}$, a pattern we call the structural–metric dissociation.
- On the Fisher-gauged empirical NTK $\tilde{\Theta}_0$, we define a reachability cost and a creation index measuring deviation from the tangent null model, and give a base-tangent reading of two recent output-space laws of post-training.

The base tangent J_0 enters through two contractions. The empirical NTK $\Theta_0 = J_0 J_0^\top$ governs output-space reachability, transfer, forgetting, and the creation residual. Circuit localization and sign are carried by a componentwise read of the same tangent, the direct logit attribution (DLA), derived from J_0 in Section 3.2 (raw parameter-gradient norms are also contractions of J_0 but fail as localizers).

2 Preliminaries

Let $f(x;\theta) \in \mathbb{R}^C$ be logits, $\pi_\theta = \mathrm{softmax}\, f$, and $\Delta\theta = \theta - \theta_0$.

Assumption 1 (Linearization / lazy regime (A1)). *$f(x;\theta_0+\Delta\theta) \approx f(x;\theta_0) + J_0(x)\Delta\theta$, with $J_0(x) = \partial_\theta f(x;\theta)\big|_{\theta_0}$ and eNTK $\Theta_0(x,x') = J_0(x)J_0(x')^\top$.*

Under Assumption 1 the only object needed is J_0, Θ_0, obtained from one forward and backward pass on the base model. The categorical (softmax) loss curvature in logit space is the Fisher $H = \mathrm{diag}(p) - pp^\top$ with $\ker H = \mathrm{span}(\mathbf{1})$. We measure reachability and forgetting in the Fisher gauge $\tilde{v} = H^{1/2}v$, $\tilde{\Theta}_0 = H^{1/2}\Theta_0 H^{1/2}$.

3 Base-Tangent Forecasts

3.1 Three A Priori Predictions

Localization. The GD update is $\Delta\theta = \eta J_0^\top \delta$ with output-space error δ ($\delta = y - p$ for SFT). Restricting to a circuit c by a projector P_c, $\|P_c\Delta\theta\| = \eta\|P_c J_0^\top \delta\|$ governs raw parameter movement. Because J_0 is fixed at θ_0 under Assumption 1, the recruited subnetwork is largely independent of task and seed, matching the cross-seed and cross-dataset overlap of Mukherjee et al. (2025). Raw parameter movement, however, is dominated by activation scale and depth and does not localize onto the interpretable circuit (Section 4.1). The DLA of Section 3.2 does, and we use it as the primary a priori score.

Reachability. The min-norm update reaching a desired logit change v has cost $\mathrm{Cost}(v) = \tilde{v}^\top \tilde{\Theta}_0^{\,+}\tilde{v} = \sum_{k:\lambda_k>0}(u_k^\top \tilde{v})^2/\lambda_k$: large-eigenvalue directions are cheap (elicitation), zero-eigenvalue directions are unreachable (creation-resistant). A second, softmax-specific source of creation-resistance appears in the Fisher gauge: moving mass toward a near-zero-probability token forces $\tilde{v}$ into the near-null of H, exploding the cost.

Forgetting. A new-task update perturbs an old capability (x_{old}, u) by $u^\top J_0(x_{\mathrm{old}})\Delta\theta$, governed by the cross-task eNTK $\Theta_{\mathrm{old,new}} = J_0(x_{\mathrm{old}})J_0(x_{\mathrm{new}})^\top$.

3.2 Deriving the Sign Forecast

The sign forecast follows from the tangent account under alignment conditions we measure rather than assume. Fix $a_h(\theta) = \langle \mathrm{LN}_0(z_{l,h}(x;\theta)W_O^{(l,h)}), v\rangle$ (final-LN scale detached: a no-gradient gauge at the evaluated checkpoint), so per-head attributions decompose the task logit exactly and move only when head h's write does, not when the model's output scale does (every gauge used below is mapped in Table 2).

Lemma 1 (ΔDLA is a contraction of the base tangent). *Under Assumption 1 applied componentwise (the observables a_h, like f, are first-order in $\Delta\theta$), one GD step $\Delta\theta = \eta J_0^\top\delta$ gives $\Delta\mathrm{DLA}_h = \eta\,\kappa_h(0)$ with $\kappa_h(0) = \sum_i \langle \nabla_\theta a_h(\theta_0),\, J_0(x_i)^\top \delta_i\rangle$.*

Proposition 1 (sign transport without smallness). *Along any gradient-flow path θ_t, $t\in[0,T]$, in the fixed normalization gauge, $\Delta\mathrm{DLA}_h = \int_0^T \kappa_h(t)\,dt$ with $\kappa_h(t) = \sum_i \langle \nabla_\theta a_h(\theta_t),\, J_t(x_i)^\top \delta_i(t)\rangle$; the detached-scale drift adds a residual outside κ, small in the runs studied (Appendix D). If κ_h keeps a constant sign on $[0,T]$ (**C1**, no-crossing), then $\mathrm{sign}(\Delta\mathrm{DLA}_h) = \mathrm{sign}(\kappa_h(0))$, a base-model quantity. C1 is an order condition independent of laziness, which does not imply it (Appendix D). Under C2 below all $\kappa_h(t)$ share one scalar mode, so C1 holds while the error keeps its readout sign (C0).*

Proposition 2 (kernel-sign alignment). *Suppose the training error projects on the task readout with a fixed sign s (**C0**) and the circuit heads share a dominant tangent direction, i.e. the component kernel $K_{hc} = \langle \nabla_\theta a_h, \nabla_\theta a_c\rangle$ is approximately rank-one with weights proportional to the base attributions (**C2**). Then $\mathrm{sign}(\kappa_h(0)) = s\cdot\mathrm{sign}(\mathrm{DLA}_{\mathrm{base},h})$ up to the rank-one residual (proofs in Appendix D).*

Corollary 1 (the sealed sign forecast, derived). *If C0 holds along the path and the kernel alignment of C2 persists along it (C2-persistence, Appendix D), every circuit head moves with sign $s\cdot\mathrm{sign}(\mathrm{DLA}_{\mathrm{base},h})$ for updates of any size: a task-improving update ($s=+1$) amplifies positive heads and deepens negative ones; a task-destroying update ($s=-1$) flips every sign together.*

The base-measurable parts of these conditions are verified on GPT-2 IOI: $\mathrm{sign}(\kappa_h(0)) = \mathrm{sign}(\mathrm{DLA}_{\mathrm{base}})$ on $12/12$ circuit heads, kernel top-eigenvalue share 0.76, dominant eigenvector separating Name Movers from Negative Name Movers exactly (Appendix D). C1 and C2 have independent support (Ji & Telgarsky, 2020; Saxe et al., 2014; Ba et al., 2022), and the path identity of Proposition 1 is the per-component analogue of influence path integrals (Pruthi et al., 2020). These conditions are consistent with five findings reported below: objective-invariance (the objective enters only through $\mathrm{sign}(s)$), sign survival at DPO $\mathcal{C} = 0.87$, the divergence boundary, the reversal fine-tune ($s{=}-1$), and over-drive breaking the sign at positive gain (C2 failure). Appendix D also derives the survival law itself. Under **C2-persistence** ($\kappa_h(t) = g_h\, c(t) + r_h(t)$, fixed loadings, exact in a rank-one-kernel model and, beyond laziness, in an aligned deep-linear network), $\mathrm{sign}(\Delta\mathrm{DLA}_h) = \mathrm{sign}(g_h) \cdot \mathrm{sign}(G)$ for the realized gain G, with no pointwise sign-constancy, and all heads collapse together exactly at the gain sign-flip. The derived predictor matches realized signs (held-out evaluation batch of the same runs) on $60/60$ head–run pairs.

3.3 The Reachability–Forgetting Duality

We solve the duality exactly in a linear model $f(x; W) = Wx$, where Assumption 1 holds exactly. With probe inputs X and target V, $\Theta_0 = G \otimes I_C$, $G = X^\top X$ (Figure 4).

Proposition 3 (Min-norm update and cost)**.** *$W^\star = V_{\mathrm{new}} G_{\mathrm{new}}^{+} X_{\mathrm{new}}^\top$ and $\mathrm{Cost}(V_{\mathrm{new}}) = \mathrm{tr}(G_{\mathrm{new}}^{+} V_{\mathrm{new}}^\top V_{\mathrm{new}}) = \|W^\star\|_F^2$.*

Theorem 1 (Duality)**.** *The same $W^\star$ perturbs old inputs only through their projection onto the new-task input subspace: $\Delta f(x_{\mathrm{old}}) = W^\star P_{\mathrm{new}} x_{\mathrm{old}}$, where $P_{\mathrm{new}} = X_{\mathrm{new}} G_{\mathrm{new}}^{+} X_{\mathrm{new}}^\top$. Equivalently $\Delta f(X_{\mathrm{old}}) = V_{\mathrm{new}}\, G_{\mathrm{new}}^{+} G_{\mathrm{new,old}}$. Orthogonal old inputs are preserved exactly.*

Corollary 2 (Sign law)**.** *For an old capability "preserve u at x_{old}," $\delta(\mathrm{old}) = (u^\top V_{\mathrm{new}})\,(G_{\mathrm{new}}^{+} X_{\mathrm{new}}^\top x_{\mathrm{old}})$. For a single new probe, $\delta(\mathrm{old}) = (u^\top v_{\mathrm{new}})\,(x_{\mathrm{new}}^\top x_{\mathrm{old}})/\|x_{\mathrm{new}}\|^2$, so $\mathrm{sign}\,\delta = \mathrm{sign}($output alignment$) \cdot \mathrm{sign}($input similarity$)$: positive transfer when both agree, interference (forgetting) when they oppose.*

Unlike a scalar forgetting bound, this predicts which held-out capability improves or degrades. On real GPT-2 the base tangent, contracted against the realized update (one JVP, deviation (iv) in Appendix E), signs an IOI fine-tune's effect on a battery of held-out capabilities with 0.88 agreement (corr. 0.83, 8 seeds, Section 4.1), which Zhang et al. (2026) explain only post hoc via data influence. All of Section 3.3 is verified to $\max|\mathrm{err}| \sim 10^{-16}$ against brute-force least squares (Table 5). The same operator supplies a base-tangent microfoundation for two output-space laws, the Invisible Leash's support-preservation premise and RL's Razor's KL geometry (Appendix C).

3.4 The Tangent Null Model and the Creation Index

The loss Hessian splits exactly as $\nabla_\theta^2 L = J^\top H J + \mathcal{E}$, with $\mathcal{E} = \sum_c (\partial_{f_c} \ell) \nabla_\theta^2 f_c$ (Schraudolph, 2002; Martens, 2020) the functional Hessian (Singh et al., 2021). Under Assumption 1 $\mathcal{E} \equiv 0$ and the GGN-Laplace is exact (Daxberger et al., 2021). We turn this into a measurable function-space quantity. Define the residual $\mathcal{R}(x) = \Delta f(x) - J_0(x)\Delta\theta$, the realized change the base tangent does not explain, and the bounded creation index $\mathcal{C} = \sum_x \|\mathcal{R}(x)\|_H^2 / \big(\sum_x \|\mathcal{R}(x)\|_H^2 + \sum_x \|J_0(x)\Delta\theta\|_H^2\big) \in [0, 1]$. By construction $\mathcal{C}$ measures the deviation of the realized update from the tangent null model: $\mathcal{C} \approx 0$ when the update is tangent-reachable and $\mathcal{C} \to 1$ when it is not. We interpret low $\mathcal{C}$ as elicitation and high $\mathcal{C}$ as capability creation, but the results below use $\mathcal{C}$ only as a tangent-deviation metric.

Proposition 4 (Null model)**.** *Under Assumption 1, $\mathcal{C} = 0$ identically; $\mathcal{C} > 0$ iff the update left the lazy regime. Significance is assessed against a null of $\mathcal{C}$ from random targets at matched $\|\Delta\theta\|$ (creation z-score).*

Proposition 5 (Separation: lazy cannot, rich can)**.** *There exist targets in $\ker\Theta_0$ (reachability cost $= \infty$) that linearized fine-tuning cannot reach for any $\Delta\theta$, yet genuine (feature-learning) fine-tuning reaches, with $\mathcal{C} > 0$. More generally, on hard (small-eigenvalue) eigendirections the tangent overprices the capability: the realized $\|\Delta\theta\|^2 \ll v^\top \tilde{\Theta}_0^{\,+} v$, and the gap is the creation signal. We verify both the exact (cost $= \infty$, linearized loss floor 12.0 vs. nonlinear $\to 0$) and graded (16–$21\times$ over-pricing, $\mathcal{C} : 5\times10^{-5} \to 0.59$) cases on CPU.*

$\mathcal{C}$ offers a tangent-deviation expression of the dichotomy between elicitation and creation (Yue et al., 2025; Li & Liu, 2026; Donoway et al., 2026). We track eNTK drift by kernel–target alignment, not

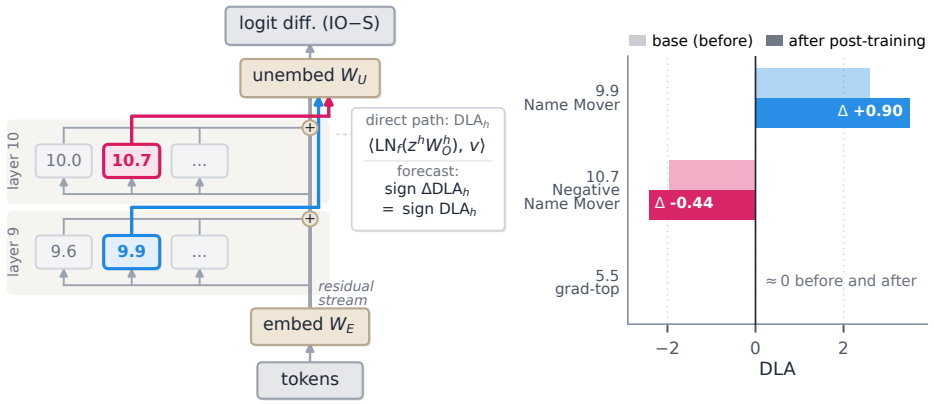

Figure 2: **Direct logit attribution and the sign forecast.** *Left:* heads read from and add to the residual stream, and the direct path of a head through W_U to the IO−S logit difference defines its DLA, with $v = W_U[:, \mathrm{IO}] - W_U[:, \mathrm{S}]$. *Right:* base DLA and DLA after SFT for a Name Mover, a Negative Name Mover, and the naive-gradient top head (plotted seed).

raw norm, following Atanasov et al. (2022). Kernel-distance and velocity diagnostics of eNTK evolution predate our index (Fort et al., 2020).

4 Experiments

Setup. We study ten models from four families (124M–1.5B) on the IOI (Wang et al., 2023), Greater-Than (Hanna et al., 2023), and Successor (Gould et al., 2024) circuits. The a priori head score is the base direct logit attribution (DLA) $a_{(l,h)} = |\mathrm{DLA}_{(l,h)}|$ with sign $\mathrm{sign}(\mathrm{DLA})$ (the raw parameter-gradient $\|P_{(l,h)} J_0^\top \delta\|$ is a baseline that does not localize). Fine-tuning uses SFT (NLL), DPO, and RLVR. Full models, hyperparameters, metrics, the null model, and a map of which quantities are available before training (Table 1) are in Appendix A.

4.1 Preregistered IOI Test

We seal predictions on the IOI circuit of GPT-2 small (Wang et al., 2023) before fine-tuning. The naive tangent score $\|P_{(l,h)} J_0^\top \delta\|$ does not localize onto the IOI circuit. It is confounded by activation scale and depth and ranks induction and layer-0 heads on top (the Name Movers $\{9.9, 9.6, 10.0\}$ fall to ranks $\{77, 85, 117\}$ of 144). The score that localizes is the base direct logit attribution (DLA) (Elhage et al., 2021), the direct contribution of a head's residual-stream output to the IO−S logit difference (Figure 2). Writing $z_{l,h}(x)$ for the output of head (l, h) at the final position and $W_O^{(l,h)}$ for its output projection, $\mathrm{DLA}_{l,h}(x) = \langle \mathrm{LN}_f(z_{l,h}(x)\, W_O^{(l,h)}),\, v \rangle$ with $v = W_U[:, \mathrm{IO}] - W_U[:, \mathrm{S}]$ and LN_f the final layer norm, averaged over prompts. Computed a priori on the base model, it ranks the Name Movers at mean rank 2.1 ± 1.8 of 144 (Negative Name Movers 2.0 ± 1.1, 5 prompt seeds) versus 93 ± 17 for the naive baseline, with the correct signs (+ for Name Movers, − for Negative Name Movers). The scale-normalized gradient still ranks the circuit at $102/144$ ($\rho = -0.61$), while attribution patching (AtP; Syed et al., 2024; Kramár et al., 2024), an independent counterfactual localizer, ranks the same heads at $2.0/144$ and agrees with DLA (ρ=0.43). Gradient scores fail (rank $\sim$92–102) and output-based scores work: output magnitude (rank 5.6) and a random readout (6.4) also localize but are unsigned as deployed, and DLA on the IO−S logit difference is sharpest (2.0). The sign forecast is not a DLA-coordinate artifact either: signed AtP assigns the identical sign to every scored head ($5/5$ circuit, $10/10$ top-10, 66% over all 144) and scores 1.00 on all eight confirmatory seeds, and both localizers fall to near-chance off-circuit.

With the DLA score $a_{(l,h)} = |\mathrm{DLA}_{(l,h)}|$ and predicted sign $\mathrm{sign}(\mathrm{DLA}_{(l,h)})$ sealed, we register: **H1b** the top-5 fine-tune-modified heads (by $|\Delta\mathrm{DLA}|$) include ≥ 2 Name Movers; **H2** the Negative Name Movers change DLA with sign opposite to the Name Movers; **H4** the signed duality of Corollary 2 predicts which held-out task improves or degrades ($\geq 75\%$ sign agreement); **H5** ablating the predicted compliant head removes $\geq 50\%$ of the IOI gain.

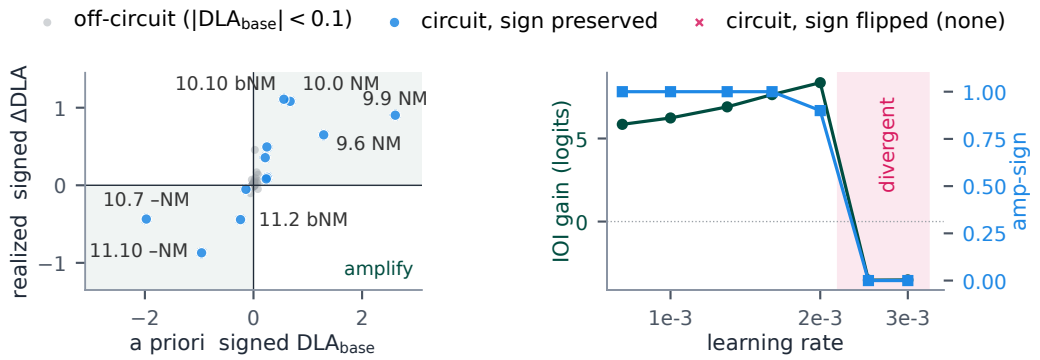

Figure 3: **Per-head prediction and confirmation on GPT-2 IOI.** *Left:* a priori $\mathrm{DLA}_{\mathrm{base}}$ vs. realized $\Delta\mathrm{DLA}$ per head (ρ=0.69, one seed). All 12 circuit heads lie on the amplify diagonal. *Right:* IOI gain (dark green, left axis) and amp-sign (blue, right axis) across learning rates. Amp-sign drops from ≥ 0.9 to 0 exactly where the gain turns negative.

On a second circuit (Greater-Than; Hanna et al., 2023) the same DLA score localizes onto the known mid-late heads ($9.1, 8.11, 8.8, 7.10, \ldots$, with 4 of the canonical heads in the top-12 vs. 1 for the naive baseline, which again returns layer-0), with base competence $+4.28$. On a third mechanism, the Successor task (increment a sequence; successor heads, Gould et al., 2024), the top-10 heads of the DLA score all lie in mid-to-late layers ($10/10$ at $L \geq 5$) while the naive gradient returns layer-0 heads ($0/10$), with base competence $+4.53$.

Mean-ablating the heads chosen by the a priori DLA score, without fine-tuning (Figure 6, Appendix B), confirms that it identifies the causal circuit and that its sign predicts the direction of effect. Ablating the predicted $-$DLA heads (Negative Name Movers) raises the IOI logit difference by $+105\%$ (they suppress IO, the copy-suppression motif of McDougall et al., 2024), while ablating the predicted $+$DLA heads removes up to 74% of it. Small $+$DLA ablations are masked by backup-name-mover self-repair (Wang et al., 2023; McGrath et al., 2023; Rushing & Nanda, 2024) (top-5: -2%, top-8: 33%, top-20: 74% of the base logit difference), which our estimator reproduces. The sealed H5 fails: one head removes $\sim 25\%$ of the fine-tune gain, below the $\geq 50\%$ threshold. The predicted set removes 74% of the base difference, and we attribute the gap to self-repair in the redundant circuit. The preregistered hypotheses and thresholds are listed in Appendix E.

Fine-tuning GPT-2 small on the IOI objective (NLL, 8 seeds, IOI improves by $+4.8 \pm 0.15$ logits) confirms the forecast (Table 4, Appendix B). The five most-changed heads (by realized $|\Delta\mathrm{DLA}|$) are the five predicted circuit heads in all 8 seeds (three Name Movers, a Negative Name Mover, and a backup Name Mover, H1b), the realized sign matches on every circuit head in every seed (H2, amplify-sign 1.00 ± 0.00), and $\rho = 0.65 \pm 0.04$ (Figure 3, left). The same pattern holds on Greater-Than ($+5.1$, 5 seeds: $\rho = 0.85$, amp 0.80, overlap $4.6/10$) and Successor ($+6.0$, 5 seeds: $\rho = 0.77$, amp 0.92, overlap $8.8/10$, $\mathcal{C} = 0.07$, elicitation). Which heads move matches the base circuit, and the direction they move matches the sign of the task error. A sealed reversal fine-tune tests this (Corollary 1, $s = -1$). Training toward the subject token at the same small learning rate (8 seeds, $\mathcal{C} = 0.28$, no divergence) flips the realized sign on $9/10$ predicted heads in every seed while the top-changed set stays the predicted circuit, as the sign-negated sealed forecast predicts. The amplify-existing-circuit alternative scores 0.10 on the same heads. The creation index of the fine-tune (Section 3.4), measured by a Jacobian-vector product, is low ($\mathcal{C} = 0.14$). The update is elicitation-dominated and stays in the tangent-reachable regime, where the base attribution is expected to forecast the realized change.

The sign forecast generalizes across ten models from four families (GPT-2 sm/med/lg/xl, Pythia-410m/1b/1.4b, Qwen2.5-0.5b/1.5b, GPT-Neo-125m; up to 1.5B). Agreement ranges 0.60–1.00, strongest where base $|\mathrm{DLA}|$ is most concentrated (GPT-2 small, Pythia-1b, Qwen2.5-1.5b: ≥ 0.97) and weakest on two small models with distributed circuits (Qwen2.5-0.5b 0.60 mean including one configuration-failure seed, median 0.90; Pythia-410m 0.77). A pre-specified a priori crispness score (the top-5 share of base $|\mathrm{DLA}|$) tracks this only weakly (ρ=0.31, n=10) and is confounded by head count. A head-count-invariant statistic (the Gini of base $|\mathrm{DLA}|$, exploratory) gives ρ=0.69 (p=0.03,

and 0.63 at $n{=}13$ with the scale spot-checks). The base model thus partially forecasts where its own forecast is weak, a tendency rather than an established law. The sign forecast also holds on base models up to 7B and on an instruction-tuned base (0.77–1.00). The single failure (Appendix B) is a 7B-instruct run whose SFT barely moves the task (gain $+1.4$), with the sign at chance while localization holds ($\rho{=}0.71$).

For H3 (reachability), within the same IOI setup, raising the in-support IO token is elicitation ($\mathcal{C} = 0.03$), whereas forcing an out-of-support third name (base probability $\sim 2 \times 10^{-3}$) requires $23\times$ more creation ($\mathcal{C} = 0.76$). This is the creation resistance under the softmax probability gauge (Section 3.4), measured on the real model.

To test whether the DLA forecast merely rediscovers published circuit annotations, we use a fourth task with no imported head list: in-context key–value lookup (single-token pairs, then a query key; GPT-2 small, base accuracy 0.87). The task readout alone (value $-$ foil) seals the top-10 signed-DLA heads. Ablating the sealed $+$DLA heads removes 68% of the lookup logit difference at $k{=}5$ (82% at $k{=}8$) and ablating the $-$DLA heads raises it, while gradient, output-magnitude, and random localizers each remove $\leq 7\%$, a $9.6\times$ margin over the strongest baseline. After fine-tuning, realized $|\Delta\mathrm{DLA}|$ concentrates on the sealed heads (top-10 overlap $6.4/10$, 10 seeds) with the sealed signs (amplify-sign 1.00 ± 0.00) despite a creation-heavy update ($\mathcal{C} = 0.87$). Ablating the base-sealed heads in the fine-tuned model removes 52% of its competence vs. $\leq 21\%$ for the three baseline localizers. The control holds on GPT-2 medium and large and Pythia-410m (3 seeds each: sign ≥ 0.97, overlap 5–$8/10$), where the sealed heads remain the strongest a priori causal localizer after fine-tuning (11–21% vs. $\leq 4\%$).

A second control, subject-verb number agreement (a computation over the head noun rather than a copy; Linzen et al., 2016), rules out an effect specific to copy or induction tasks. The blind seal again predicts the fine-tuning change (amp-sign 0.90, top-10 overlap $8.2/10$, 5 seeds), and the sealed $+$DLA heads carry 81% of the fine-tuned competence versus $\leq 4\%$ for all baselines. It also passes on GPT-2 medium (amp-sign 0.90, overlap $8.0/10$, sealed heads remove 61% vs. $\leq 5\%$).

A signed, magnitude-matched null rules out output-scale confounding. Heads matched for output scale but not for the sealed sign remove only $1\%/3\%$ on agreement and lookup, respectively, versus $81\%/51\%$ for the sealed $+$DLA heads (the signed-null replication of the 52% above), and change IOI by only -1% versus $+105\%$ for the sealed $-$DLA heads. The forecast could thus have failed by missing the sealed set, by flipped signs, or by a signed null reproducing the causal effect. None occurs on the concentrated circuits.

4.2 Zeroth-Order Localization versus First-Order Calibration

We tested the forecast across three post-training objectives (SFT with NLL, DPO (Rafailov et al., 2023), and RLVR with a verifiable reward) and ten models from four families. The predictions licensed by the linearization A1 fall into two tiers of different robustness, a structural–metric dissociation: the structural tier largely survives high-$\mathcal{C}$ updates, while the metric tier degrades.

The support of the update (which heads change) and its polarity (the sign of each change) are determined by the zeroth-order geometry of the base circuit (Corollary 1: only $\mathrm{sign}(s)$ enters, not $\|\Delta\theta\|$). Empirically both are objective-invariant (amplify-sign 1.00 across SFT, RLVR, and DPO, with the sealed top-5 criterion met in every seed of every objective) and stable across scale and family (0.60–1.00 over ten models, ≥ 0.90 on the crisp-circuit ones). GRPO-style RLVR (on-policy groups, group baseline, KL-to-base; Shao et al., 2024) preserves the forecast (1.00 and 0.90 on GPT-2 and Pythia-1b), as does $K{=}4$ generative GRPO with a verifiable reward (1.00 and 0.90), so the invariance is not a single-sample RL artifact. On the lookup and agreement controls, DPO and RLVR also preserve the sealed sign forecast (amplify-sign 0.90–1.00), including high-$\mathcal{C}$ DPO runs ($\mathcal{C} = 0.91$–0.98).

DPO leaves the lazy regime by every measure ($\mathcal{C} = 0.87$ vs. ≈ 0.15 for SFT and RLVR) yet modifies the same a priori-predicted heads with the same signs. A dense sweep (objective $\times$ lr $\times$ steps $\times\beta$, 40 runs over GPT-2 and Pythia-1b) drives $\mathcal{C}$ up to 0.86, while the top-10 sign forecast holds at 0.80–0.90 across all eight productive high-$\mathcal{C}$ runs (gains up to $+49$), versus a composition-matched permutation chance of 0.58 and 0.82 on GPT-2 and Pythia-1b (Appendix B). Chance permutes the sealed signs at fixed sign composition, so it exceeds 0.5 for sign-unbalanced circuits. Figure 5 in

Appendix B pools all 87 runs with a creation index. The forecast fails only under extreme over-drive (lr 10^{-2}, $\sim 33\times$ the SFT rate), even at moderate $\mathcal{C}$ and positive gain. In these sweeps, sign failure tracks update over-drive (only partly tracked by $\mathcal{C}$), not whether the linearization holds. Over-drive is largely predictable before training: a fixed learning-rate rule pre-classifies $83/87$ pooled runs (Appendix B). The rank ordering of change magnitudes ($\rho \approx 0.65$) is finer and closer to first order. It is robust across our real-model range ($\rho = 0.64$ for DPO at $\mathcal{C} = 0.87$) but degrades in the toy model under extreme creation (Appendix B), so we group it empirically with the first-order tier. Its ceiling (0.61–0.65) largely reflects measurement reliability: the base |DLA| ranking has test–retest $\rho = 0.69$ across disjoint evaluation batches (0.90 within the top-20, Appendix B), close to the realized $\rho \approx 0.65$. The central signal is thus the localization set (top-5 = predicted heads in every seed) and its gap to the naive-gradient reference (2.1 vs. 93 mean rank), not a scalar ρ near 1.

The first-order tier holds the calibrated predictions: exact ΔDLA magnitudes, the absolute recruitment cost $\tilde{v}^\top \tilde{\Theta}_0^+ \tilde{v}$, and the precision of the duality estimate. These rely on A1 and are expected to hold only in the elicitation regime. The creation index $\mathcal{C}$, a bounded residual ratio relative to the base-tangent prediction measured after training, separates the tiers. It partitions SFT and RLVR (elicitation, $\mathcal{C}$=0.14 and 0.16) from DPO (creation, 0.87) with non-overlapping 95% CIs under every gauge we tested (Table 3). On IOI, the low $\mathcal{C}$ of RLVR reproduces at the mechanistic level the support confinement of the Invisible Leash (Wu et al., 2025). On the lookup control both DPO and RLVR are creation-heavy, so the partition is task-dependent. In a learning-rate sweep (6 rates $\times 3$ seeds, SFT), the first-order calibration error (held-out duality precision, $1-\mathrm{corr}$) rises monotonically with $\mathcal{C}$ (Spearman $\rho(\mathcal{C}, \mathrm{CE}) = 0.79$, with CE $0.005 \to 0.82$ as $\mathcal{C}$ $0.006 \to 0.52$). The zeroth-order tier is invariant across every task-improving rate (amp-sign 1.00 up to lr 10^{-3}, and for DPO at $\mathcal{C} = 0.87$) and breaks only when the update diverges (lr 3×10^{-3}, IOI gain -3.5, circuit destroyed). The literal preregistered criterion (amp-sign ≥ 0.9 at every rate) therefore fails at the divergent extreme. The re-registered structural criterion M2′ (structural robustness for productive updates with gain>0) is confirmed out-of-sample on fresh seeds: the metric criterion M1 (calibration error rising with $\mathcal{C}$) holds ($\rho = 0.83$), and the unique divergent rate is the unique rate with amp-sign < 0.9. A fine sweep through the transition locates the boundary (Figure 3, right): amp-sign is 1.00 up to lr 1.6×10^{-3} and 0.90 at 2×10^{-3} (peak gain $+8.4$), then drops to 0 at 2.5×10^{-3}, where the gain turns negative.

The boundary is not simply gain>0: on Greater-Than an extreme DPO update (gain $+45$) breaks the sign (0.50) while SFT and RLVR preserve it (0.80 and 0.73). The hierarchy thus places the central assumption (A1) in the metric tier alone, while the structural predictions rest on the geometry of the base model. We therefore do not claim that global NTK linearization explains post-training. The base tangent serves as a predictive diagnostic: the structural tier is robust beyond the lazy regime, and the metric tier is diagnosed by $\mathcal{C}$ and may fail.

5 Related Work

Output-space accounts of post-training hold that fine-tuning largely elicits capabilities already present in the base model (Yue et al., 2025; Karan & Du, 2026). Wu et al. (2025) argue that RLVR largely stays within the base support and rarely exceeds its pass@k, and Shenfeld et al. (2026) show that forgetting is governed by the KL to the base policy, which on-policy RL implicitly minimizes. These distributional laws are measured during or after training and predict neither which internal components change nor which capabilities are reachable. We relate the support-preservation premise of the Invisible Leash and RL's Razor to one base-model object, the base eNTK $\Theta_0 = J_0 J_0^\top$. Its reachable cone bounds the support change behind the Leash (we address this premise, not the full pass@k ceiling), and its Fisher gauge $F = J_0^\top H J_0$ sets the KL norm $\frac{1}{2}\Delta^\top F \Delta$ that on-policy RL minimizes. It also predicts localization and signed transfer.

Tangent-space and NTK analyses of fine-tuning provide our methodological basis. Base-NTK overlap predicts forgetting magnitude (Doan et al., 2021), and closed-form analyses give task-level signed laws for forgetting and transfer (Lin et al., 2023; Wan et al., 2025; Goldfarb et al., 2024). Concurrently, Hidekel & Raviv (2026) derive a closed-form base-NTK predictor of the forgetting vector before any gradient step, but in output space for continual adaptation, whereas ours is componentwise on circuits and causally tested. Other work edits models via task vectors (Ortiz-Jimenez et al., 2023; Li et al., 2025a), gives a kernel view of fine-tuning (Malladi et al., 2023), or predicts as-

sociation formation in pretraining (Im et al., 2026). Rahimi Afzal et al. (2025) show that fine-tuning is well approximated by its linearization but read out a single scalar whole-model risk, whereas we read it per component and signed. Zhang et al. (2026) study the reduced forgetting of RFT post hoc via data influence, and REN & Sutherland (2025) predict the DPO squeezing effect in output space rather than on circuits. We instead forecast, before training, which circuit components an SFT, RL, or DPO update changes and the sign of each change, and show that the structural predictions hold even where the linearization fails.

Mechanistic studies of fine-tuning-induced circuit change measure the damage after the fact: head-level differential vulnerability, with RL preserving circuits better than SFT (Nunez et al., 2026); circuit-aware fine-tuning (Wang et al., 2025); IOI-specific corruption (Chhabra et al., 2025); and head disruption in continual fine-tuning (Lundstrom-Imanov, 2026). Post-hoc skill localization (Panigrahi et al., 2023), update sparsity (Mukherjee et al., 2025), and latent-direction reuse (Ward et al., 2025) are likewise retrospective, and causal localization can mispredict editing sites (Hase et al., 2023). Mechanistic model diffing (Lindsey et al., 2024; Minder et al., 2025; Shportko et al., 2026) isolates what tuning added from paired checkpoints, whereas our forecast needs only the base model. Tigges et al. (2024) track circuit algorithms that stay consistent across training and scale while components turn over beneath them. Prakash et al. (2024) show after training that fine-tuning amplifies an existing circuit, Jain et al. (2024) find that it often learns a minimal wrapper over base capabilities, Li et al. (2025b) frame it as subgraph search and observe head-level sign reversals, and Prakash et al. (2025) improve reasoning by updating only circuit components identified post hoc. These works identify or modify circuits during or after training, whereas we seal a signed forecast before training and test it causally.

Work on capability creation frames it via support change or description length (Li & Liu, 2026; Donoway et al., 2026), and the empirical NTK has served as a post-hoc interpretability lens (Lin, 2025). We instead measure the distinction as a function-space residual and use the base eNTK predictively. The Three-Gate account (Zhu et al., 2025) shows post hoc that RLVR lands in low-curvature subspaces, the bias our base eNTK encodes a priori, consistent with near-linear RLVR trajectories (Wang et al., 2026; Tomihari, 2026). Linearized attention does not reach the kernel regime at practical widths (Miñoza et al., 2026). Predictive, falsifiable interpretability remains an open problem (Leavitt & Morcos, 2020; Sharkey et al., 2025), which we operationalize as sealed preregistration (Forde & Paganini, 2019).

6 Conclusion

In the settings we study, the base tangent J_0 forecasts, before training, which heads an update will change and in which direction. The sign forecast holds beyond the linearization that motivates it, while magnitude forecasts require the lazy regime, a structural–metric dissociation. With causal ablation, sealed forecasts allow circuit-level effects of post-training to be tested against predictions made before training.

Limitations. We validate on three established circuits, which give independent ground truth for sealed per-head prediction, plus the lookup and agreement controls with no imported head list (Section 4.1). Large-scale RLHF, reasoning, and safety alignment remain future work. The constraint is methodological rather than one of scale (four families up to 7B, with SFT, DPO, RLVR, and GRPO). Metric forecasts hold in elicitation ($\mathcal{C} \approx 0.14$) and degrade under feature learning, while the structural forecast survives creation-heavy updates (Section 4.2). The boundary is derived only in tractable models (Appendix D). RL evidence spans one-token RLVR and $K{=}4$ GRPO, attribution beyond $\sim$7B is memory-bound, and Pythia-410m is weaker ($5/10$ vs. $8/10$, in line with the crispness tendency of Section 4.1). Metrics use 3–10 seeds. GPT-2 small with IOI and SFT is the original sealed setting, and a subsequent prospective sealed battery over held-out distributions and models shows the same pattern: pre-declared positive cells pass, and a pre-declared weak cell is weakest (Appendix E). The remaining analyses were planned in advance but not sealed, hence exploratory, and head indices must be re-locked for each new model.

REPRODUCIBILITY STATEMENT

All analytic claims (Sections 3.3–3.4) are verified by numpy and torch scripts that run on CPU in seconds, and the base-model a priori scores (the sealed predictions) are likewise computable on CPU. Fine-tuning uses a single GPU. Appendix A gives the models, data, hyperparameters, and metrics. Appendix E lists the preregistered hypotheses, their acceptance thresholds, and every deviation from them, and Table 7 reports seed-bootstrap confidence intervals for the primary numbers.

The sealing protocol is fully specified in Appendix E, so each new behavioral direction (e.g. refusal, Arditi et al., 2024) can be tested in the same preregistered form.

AI USE STATEMENT

Large language models were used as a research and writing assistant throughout this project: retrieving and summarizing related work, drafting and refactoring analysis code, executing the preregistered verification protocols, and editing prose. All quantitative claims were verified against the frozen result files, the sealed predictions, and the CPU verification suite, and the authors reviewed and take full responsibility for all content.

REFERENCES

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, volume 37, pp. 136037–136083, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/f545448535dfde4f9786555403ab7c49-Paper-Conference.pdf.

Alexander Atanasov, Blake Bordelon, and Cengiz Pehlevan. Neural Networks as Kernel Learners: The Silent Alignment Effect. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=1NvflqAdoom.

Jimmy Ba, Murat Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 37932–37946, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/f7e7fabd73b3df96c54a320862afcb78-Paper-Conference.pdf.

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. URL https://aclanthology.org/2025.findings-naacl.170/.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/ae614c557843b1df326cb29c57225459-Paper.pdf.

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux - effortless bayesian deep learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 20089–20103, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/a7c9585703d275249f30a088cebba0ad-Paper.pdf.

Thang Doan, Mehdi Abbana Bennani, Bogdan Mazoure, Guillaume Rabusseau, and Pierre Alquier. A theoretical analysis of catastrophic forgetting through the NTK overlap matrix. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1072–1080, 2021. URL https://proceedings.mlr.press/v130/doan21a.html.

Elizabeth Donoway, Hailey Joren, Fabien Roger, and Jan Leike. Excess description length of learning generalizable predictors. *arXiv preprint arXiv:2601.04728*, 2026.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Jessica Zosa Forde and Michela Paganini. The scientific method in the science of machine learning. *arXiv preprint arXiv:1904.10922*, 2019.

Stanislav Fort, Gintare Karolina Dziugaite, Mansheej Paul, Sepideh Kharaghani, Daniel M Roy, and Surya Ganguli. Deep learning versus kernel learning: an empirical study of loss landscape geometry and the time evolution of the neural tangent kernel. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5850–5861, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/405075699f065e43581f27d67bb68478-Paper.pdf.

Daniel Goldfarb, Itay Evron, Nir Weinberger, Daniel Soudry, and Paul Hand. The joint effect of task similarity and overparameterization on catastrophic forgetting —an analytical model. In *International Conference on Learning Representations*, volume 2024, pp. 25160–25245, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/6cb00ce1a21a090a3dae04cebebd8341-Paper-Conference.pdf.

Rhys Gould, Euan Ong, George Ogden, and Arthur Conmy. Successor heads: Recurring, interpretable attention heads in the wild. In *International Conference on Learning Representations*, volume 2024, pp. 9236–9261, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/2722a0ccf6acfe3d144fdbb0dedd80b5-Paper-Conference.pdf.

Michael Hanna, Ollie Liu, and Alexandre Variengien. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 76033–76060, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/efbba7719cc5172d175240f24be11280-Paper-Conference.pdf.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 17643–17668, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/3927bbdcf0e8d1fa8aa23c26f358a281-Paper-Conference.pdf.

Ido Nitzan Hidekel and Dan Raviv. Catastrophic forgetting is low-rank: A function-space theory for continual adaptation. *arXiv preprint arXiv:2606.18024*, 2026.

Shawn Im, Changdae Oh, Zhen Fang, and Sharon Li. How do transformers learn to associate tokens: Gradient leading terms bring mechanistic interpretability. In *International Conference on Learning Representations*, volume 2026, pp. 156381–156419, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/fd4e8712bc4aaa97cc51cf93ffe2366d-Paper-Conference.pdf.

Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/5a4be1fa34e62bb8a6ec6b91d2462f5a-Paper.pdf.

Samyak Jain, Robert Kirk, Ekdeep Singh Lubana, Robert Dick, Hidenori Tanaka, Tim Rocktaeschel, Edward Grefenstette, and David Krueger. Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks. In *International Conference on Learning Representations*, volume 2024, pp. 9944–10015, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/29b5de0a8925cb4920d9c64542108335-Paper-Conference.pdf.

Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 17176–17186, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/c76e4b2fa54f8506719a5c0dc14c2eb9-Paper.pdf.

Aayush Karan and Yilun Du. Reasoning with sampling: Your base model is smarter than you think. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Vsgq2ldr4K.

János Kramár, Tom Lieberum, Rohin Shah, and Neel Nanda. AtP*: An efficient and scalable method for localizing LLM behaviour to components. *arXiv preprint arXiv:2403.00745*, 2024.

Matthew L. Leavitt and Ari Morcos. Towards falsifiable interpretability research. *arXiv preprint arXiv:2010.12016*, 2020.

Hongkang Li, Yihua Zhang, Shuai Zhang, Meng Wang, Sijia Liu, and Pin-Yu Chen. When is task vector provably effective for model editing? a generalization analysis of nonlinear transformers. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=vRvVVb0NAz.

Yueyan Li, Wenhao Gao, Caixia Yuan, and Xiaojie Wang. Fine-tuning is subgraph search: A new lens on learning dynamics. *arXiv preprint arXiv:2502.06106*, 2025b.

Yuhao Li and Shengchao Liu. On distinguishing capability elicitation from capability creation in post-training: A free-energy perspective. *arXiv preprint arXiv:2605.08368*, 2026.

Jennifer Lin. Feature identification via the empirical NTK. *arXiv preprint arXiv:2510.00468*, 2025.

Sen Lin, Peizhong Ju, Yingbin Liang, and Ness Shroff. Theory on forgetting and generalization of continual learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 21078–21100, 2023. URL https://proceedings.mlr.press/v202/lin23f.html.

Jack Lindsey, Adly Templeton, Jonathan Marcus, Thomas Conerly, Joshua Batson, and Christopher Olah. Sparse crosscoders for cross-layer features and model diffing. Transformer Circuits Thread, 2024. https://transformer-circuits.pub/2024/crosscoders/index.html.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535, 2016. URL https://aclanthology.org/Q16-1037/.

Gustav Olaf Yunus Laitinen-Fredriksson Lundstrom-Imanov. Mechanistic analysis of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2601.18699*, 2026.

Sadhika Malladi, Alexander Wettig, Dingli Yu, Danqi Chen, and Sanjeev Arora. A kernel-based view of language model fine-tuning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 23610–23641, 2023. URL https://proceedings.mlr.press/v202/malladi23a.html.

James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020. URL http://jmlr.org/papers/v21/17-678.html.

Callum Stuart McDougall, Arthur Conmy, Cody Rushing, Thomas McGrath, and Neel Nanda. Copy suppression: Comprehensively understanding a motif in language model attention heads. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 337–363, 2024. URL https://aclanthology.org/2024.blackboxnlp-1.22/.

Thomas McGrath, Matthew Rahtz, János Kramár, Vladimir Mikulik, and Shane Legg. The hydra effect: Emergent self-repair in language model computations. *arXiv preprint arXiv:2307.15771*, 2023.

Julian Minder, Clément Dumas, Caden Juang, Bilal Chughtai, and Neel Nanda. Overcoming sparsity artifacts in crosscoders to interpret chat-tuning. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 106423–106474, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/9902a53031ebbbab73898028073d4790-Paper-Conference.pdf.

Jose Marie Antonio Miñoza, Paulo Mario P. Medina, and Sebastian C. Ibañez. Linearized attention cannot enter the kernel regime at any practical width. *arXiv preprint arXiv:2603.13085*, 2026.

Sagnik Mukherjee, Lifan Yuan, Dilek Hakkani-Tur, and Hao Peng. Reinforcement learning finetunes small subnetworks in large language models. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 132119–132138, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/bf235a1d6780afd979f2f81676f43413-Paper-Conference.pdf.

Neel Nanda and Joseph Bloom. Transformerlens. https://github.com/TransformerLensOrg/TransformerLens, 2022.

Jeanmely Rojas Nunez, Viraj Sawant, Nathan Allen, Nomgondalai Amgalanbaatar, Yannis Zongo, Vasu Sharma, and Maheep Chaudhary. Mechanistic origins of catastrophic forgetting: Why RL preserves circuits better than SFT? *arXiv preprint arXiv:2605.28860*, 2026.

Guillermo Ortiz-Jimenez, Alessandro Favero, and Pascal Frossard. Task arithmetic in the tangent space: Improved editing of pre-trained models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 66727–66754, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/d28077e5ff52034cd35b4aa15320caea-Paper-Conference.pdf.

Abhishek Panigrahi, Nikunj Saunshi, Haoyu Zhao, and Sanjeev Arora. Task-specific skill localization in fine-tuned language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 27011–27033, 2023. URL https://proceedings.mlr.press/v202/panigrahi23a.html.

Nikhil Prakash, Tamar Rott Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. Fine-tuning enhances existing mechanisms: A case study on entity tracking. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=8sKcAWOf2D.

Nikhil Prakash, Donghao Ren, Dominik Moritz, and Yannick Assogba. Constructive circuit amplification: Improving math reasoning in LLMs via targeted sub-network updates. *arXiv preprint arXiv:2512.16914*, 2025.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. In *Advances in Neural Information Processing Systems*, volume 33, pp. 19920–19930, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/e6385d39ec9394f2f3a354d9d2b88eec-Paper.pdf.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a85b405ed65c6477a4fe8302b5e06ce7-Paper-Conference.pdf.

Zahra Rahimi Afzal, Tara Esmaeilbeig, Mojtaba Soltanalian, and Mesrob I Ohannessian. Linearization explains fine-tuning in large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=tdwRIP6NG2.

YI REN and Danica Sutherland. Learning dynamics of LLM finetuning. In *International Conference on Learning Representations*, volume 2025, pp. 70523–70563, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/afe1aa79e5eea7955f553c61a307273e-Paper-Conference.pdf.

Cody Rushing and Neel Nanda. Explorations of self-repair in language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 42836–42855, 2024. URL https://proceedings.mlr.press/v235/rushing24a.html.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations*, 2014. URL http://arxiv.org/abs/1312.6120.

Nicol N. Schraudolph. Fast curvature matrix-vector products for second-order gradient descent. *Neural Computation*, 14(7):1723–1738, 2002.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y.K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeffrey Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Isaac Bloom, Stella Biderman, Adrià Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Mary Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, William Saunders, Eric J Michaud, Stephen Casper, Max Tegmark, David Bau, Eric Todd, Atticus Geiger, Mor Geva, Jesse Hoogland, Daniel Murfet, and Thomas McGrath. Open problems in mechanistic interpretability. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=91H76m9Z94. Survey Certification.

Idan Shenfeld, Jyothish Pari, and Pulkit Agrawal. RL's razor: Why online reinforcement learning forgets less. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=7HNRYT4V44.

Andrii Shportko, Shubham Bhokare, Ahmed Zeyad A Alzahrani, Bowen Cheng, Gustavo Mercier, and Jessica Hullman. Localizing RL-induced tool use to a single crosscoder feature. In *Mechanistic Interpretability Workshop at ICML 2026*, 2026. URL https://openreview.net/forum?id=oKvgaSDa6I.

Sidak Pal Singh, Gregor Bachmann, and Thomas Hofmann. Analytic insights into structure and rank of neural network hessian maps. In *Advances in Neural Information Processing Systems*, volume 34, pp. 23914–23927, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/c900ced7451da79502d29aa37ebb7b60-Paper.pdf.

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. URL https://aclanthology.org/2024.blackboxnlp-1.25/.

Curt Tigges, Michael Hanna, Qinan Yu, and Stella Biderman. LLM circuit analyses are consistent across training and scale. In *Advances in Neural Information Processing Systems*, volume 37, pp. 40699–40731, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/47c7edadfee365b394b2a3bd416048da-Paper-Conference.pdf.

Akiyoshi Tomihari. Learning dynamics in RL post-training for language models. *arXiv preprint arXiv:2601.04670*, 2026.

Hongyi Wan, Shiyuan Ren, Wei Huang, Miao Zhang, Xiang Deng, Yixin Bao, and Liqiang Nie. Understanding the forgetting of (Replay-based) continual learning via feature learning: Angle matters. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 61956–62019, 2025. URL https://proceedings.mlr.press/v267/wan25d.html.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Tianle Wang, Jiayu Liu, Zhongyuan Wu, Shenghao Jin, Wei Chen, Hao Xu, and Ning Miao. Linear dynamics in the RLVR training of large language models. *arXiv preprint arXiv:2601.04537*, 2026.

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. URL https://proceedings.mlr.press/v267/wang25ak.html.

Jake Ward, Chuqiao Lin, Constantin Venhoff, and Neel Nanda. Reasoning-finetuning repurposes latent representations in base models. *arXiv preprint arXiv:2507.12638*, 2025.

Fang Wu, Weihao Xuan, Ximing Lu, Mingjie Liu, Yi Dong, Zaid Harchaoui, and Yejin Choi. The invisible leash: Why RLVR may or may not escape its origin. *arXiv preprint arXiv:2507.14843*, 2025.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 57654–57689, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/537d5aa768c2d534016a4d06f87bc8fb-Paper-Conference.pdf.

Zhihao Zhang, Qiaole Dong, Qi Zhang, Enyu Zhou, Jun Zhao, Zhiheng Xi, Senjie Jin, Xiaoran Fan, Yuhao Zhou, Mingqi Wu, Yanwei Fu, Tao Ji, Tao Gui, Xuanjing Huang, and Kai Chen. Why reinforcement fine-tuning enables MLLMs preserve prior knowledge better: A data perspective. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=tP01GCr0nS.

Hanqing Zhu, Zhenyu Zhang, Hanxian Huang, DiJia Su, Zechun Liu, Jiawei Zhao, Igor Fedorov, Hamed Pirsiavash, Zhizhou Sha, Jinwon Lee, David Z. Pan, Zhangyang Wang, Yuandong Tian, and Kai Sheng Tai. The path not taken: RLVR provably learns off the principals. *arXiv preprint arXiv:2511.08567*, 2025.

A EXPERIMENTAL DETAILS

Table 1: **Forecasts before training and confirmations after.** Localization, sign, and reachability use the base model alone. The creation index requires the realized update and is measured after training.

Quantity	What it captures	Forecast before training	Confirmed after
Localization	which components change	✓	✓
Sign	direction of each change	✓	✓
Reachability	cost to reach a target	✓	✓
Creation index	deviation from the base tangent	post hoc	✓

Models and tasks. Ten models from four families via TRANSFORMERLENS (Nanda & Bloom, 2022), fp32: GPT-2 {small, medium, large, xl}, Pythia {410m, 1b, 1.4b}, Qwen2.5 {0.5b, 1.5b}, and GPT-Neo-125m (124M–1.5B), on three circuits with established ground truth: **IOI** (Wang et al., 2023) (ABBA and BABA templates, readout = the IO−S logit difference at the final position) and **Greater-Than** (Hanna et al., 2023) ("...from 17XX to 17", readout = mean valid-year − mean invalid-year logit), and **Successor** (Gould et al., 2024) (increment a consecutive-integer sequence "3 4 5 6"→"7", readout = next − last-shown logit). Capabilities are restricted to single-token names, years, or integers.

Fine-tuning. Full-parameter SGD, three objectives that all increase the task capability: **SFT** with NLL (primary, lr 3×10^{-4}, 40 steps), **DPO** (preference vs. a frozen base reference, β=0.1), **RLVR** (REINFORCE, verifiable $0/1$ reward). We use 256 train and 96–128 eval prompts for IOI and Successor (seed-disjoint pools, with Greater-Than scored on its training pool) and 512 and 192 for the lookup and agreement controls (exactly disjoint for lookup, while the in-distribution agreement pools share templates), with 3–10 seeds per cell. The LR sweep uses $\{10^{-6}, 10^{-5}, 10^{-4}, 3\times10^{-4}, 10^{-3}, 3\times10^{-3}\}$.

Metrics. H1b (top-5 $|\Delta\mathrm{DLA}| \cap$ predicted heads); amp-sign (fraction of top-k base-DLA heads, k=10 unless stated, whose realized sign matches base); ρ = Spearman($|\mathrm{DLA}_{\mathrm{base}}|, |\Delta\mathrm{DLA}|$); creation index $\mathcal{C} \in [0, 1]$ (a bounded residual ratio measuring deviation from the base-tangent prediction, via a Jacobian-vector product; computed on the scalar task readout with plain L_2 norms, the Fisher-gauged full-logit form of Section 3.4 restricted to the readout direction); held-out duality check (base tangent $\times$ realized update, one JVP); recruitment cost $v^\top \tilde{\Theta}_0^+ v$. Localization reference = the naive gradient-norm ranking (predicted heads at mean rank 2.1 ± 1.8 vs. 93 ± 17, 5 seeds). Means $\pm$ std over seeds and bootstrap CIs for correlations.

Table 2: **Gauge used for each measured quantity.** Componentwise attributions use the detached-LN gauge and output-space quantities the softmax Fisher gauge. The creation index is defined on Fisher-gauged full logits and reported on the scalar readout (gap quantified in Appendix B).

Quantity	Space / gauge	Why this gauge
DLA, ΔDLA, sign	per-component, detached final-LN	isolates a head's write from global output-scale drift
κ_h, K_{hc}, loadings g_h	per-component, same detached gauge	keeps κ the exact time-derivative of the scored a_h
Creation index $\mathcal{C}$	full logits, Fisher H; reported on the readout with L_2	softmax curvature; the readout form is the restriction of the same norm
Reachability cost, duality	output space, Fisher H	$\ker H = \mathrm{span}(\mathbf{1})$ removes the logit-shift gauge freedom
eNTK drift (CKA)	output space, scale-invariant	separates kernel rotation from kernel rescaling

B Numerical verification and closed-form illustration

Scale and instruction-tuning spot checks. Beyond the sweep, the sign forecast holds at scale on base models to 7B (Pythia-2.8b/6.9b, Qwen2.5-3b/7b; amp-sign 0.77–1.00) and on an instruction-tuned base (Qwen2.5-3b-Instruct, 0.90), so it does not require a pre-RLHF model (per-model numbers in Table 6 in Appendix B). It has a single failure mode: on the 7B instruction-tuned model the SFT barely moves the task (IOI gain $+1.4$ versus $+4.8$ on GPT-2), so there is little signed change to forecast and the sign agreement falls to chance (0.40). The localization tier still holds there (ρ=0.71). Section 4.2 compares the objectives (SFT, DPO, RLVR) and shows the tier structure. DPO is creation-heavy on these circuits because off-policy training with negative examples pushes mass toward base-low-probability directions, the probability-gauge creation source of Section 3.4.

Gauge sensitivity of the creation index. $\mathcal{C}$ is defined in Section 3.4 on the full logit vector in the Fisher gauge but reported on the scalar task readout with plain L_2 norms (Appendix A), which raises two questions: does the readout restriction miss creation that lands off the readout, and does the tier separation depend on the gauge? We recompute representative runs under all four combinations of $\{$readout, full logit$\}\times\{L_2$, Fisher$\}$ from one JVP (Table 3). In these runs the separations do not depend on the gauge: the partition into elicitation (SFT, RLVR) and creation (DPO) and the in- versus out-of-support reachability separation hold with non-overlapping 95% CIs under every gauge, and the readout numbers reproduce the reported values (0.14, 0.16, 0.87, and 0.03 vs 0.76), as does the divergent learning rate (0.52). Under the formal full-logit Fisher definition the elicitation runs move down (0.05 and 0.07), so the separation widens, and the scalar readout is if anything conservative. Off-readout creation is real but does not reverse the reading: 37–41% of the Fisher residual mass of the elicitation runs lies off the readout direction (DPO 8%), yet restricting $\mathcal{C}$ to that orthogonal complement alone still separates the regimes (0.10 and 0.12 vs. 0.73). The one

gauge that compresses a separation is full-logit L_2 on reachability ($3.6\times$ instead of $23\times$), because an unweighted L_2 norm over the vocabulary ignores the probability geometry that makes an out-of-support target expensive. This is why the definition is Fisher-gauged.

Table 3: **Creation index under four gauges.** GPT-2 small, 3 seeds per cell, seed-bootstrap 95% CIs. The last row is the formal definition of Section 3.4. Sep. marks non-overlapping CIs between elicitation and creation cells. Over-drive is the divergent sweep rate (lr 3×10^{-3}, gain -3.5).

	Objective partition				H3 reachability			
Gauge	SFT	RLVR	DPO	Sep.	in-supp.	out-supp.	ratio	over-drive
readout, L_2 (reported)	0.15	0.18	0.87	✓	0.03	0.76	$23\times$	0.52
readout, Fisher	0.16	0.20	0.88	✓	0.03	0.72	$21\times$	0.52
full logit, L_2	0.23	0.27	0.85	✓	0.24	0.88	$3.6\times$	0.49
full logit, Fisher (Section 3.4)	0.05	0.07	0.81	✓	0.05	0.82	$15\times$	0.45

Table 4: **Main real-model results.** GPT-2 IOI unless noted. Brackets are seed-bootstrap 95% CIs (Table 7 in Appendix E). Stronger nulls are in Appendix B and per-model numbers in Table 6 in Appendix B.

Prediction	Test	Result
H1b localization	Name Movers in top-5 $\lvert\Delta\text{DLA}\rvert$	$3/3$, all 8 seeds
H2 sign forecast	amplify-sign	1.00 ± 0.00
Localization corr.	$\rho(\lvert\text{DLA}_{\text{base}}\rvert, \lvert\Delta\text{DLA}\rvert)$	$0.65\ [0.63, 0.68]$
Causal (intervention)	ablate $-$DLA / $+$DLA set	$+105\%$ / 74% removed
H4 duality	held-out duality check	$0.88\ [0.86, 0.91]$ (8 seeds)
H3 reachability	creation $\mathcal{C}$, in- vs out-support	$0.03\ [.03, .04]$ vs $0.76\ [.73, .80]$
Objective-invariance	amp-sign (SFT/DPO/RLVR)	1.00
Generalization	amp-sign (10 models, 4 families)	0.60–1.00
Anti-circularity	lookup, sealed from task readout only	overlap $6.4/10$; sign 1.00; causal $9.6\times$
Hierarchy (LR sweep)	M1 $\rho(\mathcal{C}, \text{CE})$; M2′ (OOS)	0.83; PASS
Prospective sealed battery	3 positive + 1 declared-weak cell	positives $3/3$; weak cell lowest

Rank-correlation ceiling (decomposition). To quantify the source of the per-head rank-correlation ceiling we decompose the squared rank residuals of the seed-0 confirmatory run (all 144 heads). The self-repair class (backup plus negative Name Movers, 10 of 144 heads) carries 0.5% of the total squared rank-residual mass versus 6.9% under a uniform spread (permutation $p = 0.9995$ for enrichment), and excluding the class lowers ρ from 0.69 to 0.62 while excluding ten random heads leaves it unchanged (0.690 ± 0.015 over 100 draws): these heads are high-magnitude, rank-conforming points. Backup recruitment is nevertheless real and stable: head 10.10, forecast rank 6, is the largest realized $\lvert\Delta\text{DLA}\rvert$ head in 15 of 16 seeds, and the backup class has median signed rank residual $+2$ versus -2 for Name Movers. The ceiling itself is a reliability floor of the ranking metric: the base $\lvert\text{DLA}\rvert$ ranking reproduces across two disjoint frozen evaluation batches at $\rho = 0.69$ (0.90 within the top-20 heads), matching the realized 16-seed mean of 0.65, so the observed ρ sits at the test-retest ceiling of the measurement rather than reflecting forecast error.

The closed-form (linear-model) results below are exact by construction: in a linear model Assumption 1 holds identically, so the duality, cost, and separation identities are algebraic and the reported $\sim 10^{-16}$ errors are consistency checks of the derivation, not empirical findings. The one row that bridges the toy to a nonlinear network is the sign-law generalization: on a small MLP trained to convergence (well outside the lazy regime), the base-tangent a priori sign predicts the realized change with 0.98 agreement versus 0.53 for the naive predictor, the same zeroth-order robustness we then confirm on transformers (Table 5). Table 6 gives the per-model sign agreement summarized in Table 4.

Boundary cases. Magnitude-matched heads also carry competence on the more distributed Greater-Than and Successor circuits, so the signed null is most diagnostic on concentrated circuits (IOI and the lookup and agreement controls, where it is clean). The standard fine-tuning configuration decreases competence on Qwen2.5-0.5B (whose RMSNorm final norm also requires different attribution handling). We treat this as a configuration- or model-specific failure rather than evidence

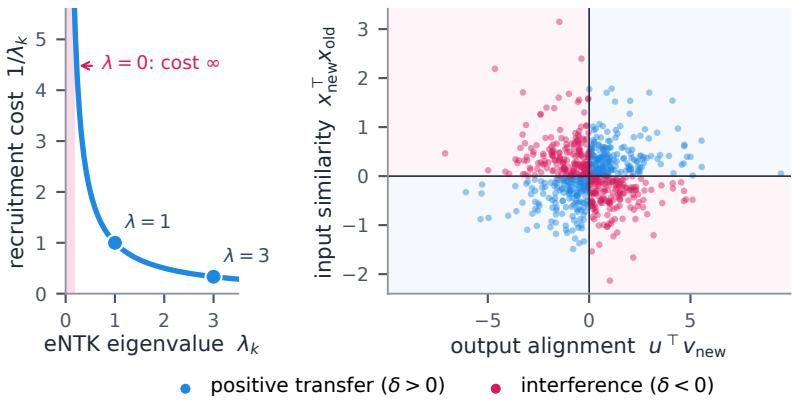

Figure 4: **Closed-form theory in the linear model.** *Left:* recruitment cost $1/\lambda_k$ of a unit target along each eNTK eigendirection, infinite for the zero-eigenvalue direction. *Right:* 700 random old and new pairs split exactly into transfer (same-sign quadrants) and interference, as the signed duality law $\operatorname{sign}\delta = \operatorname{sign}(u^\top v_{\text{new}}) \cdot \operatorname{sign}(x_{\text{new}}^\top x_{\text{old}})$ predicts.

Table 5: **Numerical checks of the theory.** Closed-form identities in the linear model are compared against direct computation, and the sign law is additionally tested on a nonlinear network.

Statement	Numerical test	Outcome
Min-norm update and cost (Prop. 3)	vs. brute-force least squares	$\max \lvert\text{err}\rvert = 4 \times 10^{-16}$
Duality and transfer (Thm. 1)	vs. direct $W^\star x_{\text{old}}$	7×10^{-16}
Sign law (Cor. 2), linear	sign agreement	exact
Sign law, nonlinear net (GD to convergence)	a priori vs. realized sign	0.98 (naive 0.53)
Forgetting bound (Section C)	2000 random draws	all hold
Null model, linear (Section 3.4)	$\mathcal{C}$ on linear f	3×10^{-32}
Creation vs. drift, nonlinear	rank corr. ($\mathcal{C}$, drift)	1.000
Separation (Prop. 5), exact	linear floor vs. nonlinear	12.0 vs. 0
Separation (Prop. 5), graded	tangent over-pricing	16–21×
RL's Razor bridge (Section C)	$\mathrm{KL}/\frac{1}{2}\Delta\theta^\top F \Delta\theta$	$\to 1.0001$
Invisible Leash (Section C)	on- vs. off-policy p_{OOS}	8×10^{-15} vs. 4×10^{-12}

against the forecast. Per-model means include failed-run seeds where they occur (Qwen2.5-0.5B: one configuration-failure seed, median 0.90). The held-out transfer forecast (H4) replicates on a second circuit: with the marginal-NLL objective used throughout, Greater-Than base-tangent sign-agreement reaches 0.99 (0.98 magnitude correlation, versus 0.52 under a mismatched raw-margin loss).

Figure 5 pools every run for which we compute a creation index (n=87: the SFT learning-rate sweep, the objective comparison, the dense high-$\mathcal{C}$ sweep, the control tasks under DPO and RLVR, and the sealed reversal). If the sign forecast were an artifact of the lazy regime, agreement would fall with $\mathcal{C}$. Instead the band stays at 0.7–1.0 across the entire range $\mathcal{C} \in [0, 0.98]$ (73 of 76 productive runs ≥ 0.8, against composition-matched permutation chance of 0.52–0.82 per cell, with the permutation null rejected jointly across the ten base models at $p < 10^{-3}$, Appendix B). Only over-driven runs collapse, defined as negative training-objective gain (raw gain > 0 for the $s = -1$ reversal) or $\mathrm{lr} \geq 10^{-2}$. The validity domain can largely be fixed in advance: a learning-rate rule fixed before evaluation (2.5×10^{-3} for unbounded-logit objectives, from the published boundary for SFT on GPT-2 IOI, and the stipulated 10^{-2} for bounded-gradient DPO and RLVR) pre-classifies 83/87 pool runs (95.4%) as productive versus over-driven using no post-training quantity. The four misses are logit-diff runs within one sweep notch of the boundary (GPT-2 over-drives at 2×10^{-3}, Pythia-1B is still productive at 3×10^{-3}), so the residual outcome-dependence is a model- and objective-dependent shift of roughly $1.5\times$ in learning rate. A gain-matched comparison further separates the two variables: binning the 76 productive runs into gain-matched quartiles, high-$\mathcal{C}$ and low-$\mathcal{C}$ runs show equal sign agreement within every bin (differences 0.00, -0.04, $+0.05$, $+0.17$), and across productive runs amp-sign is uncorrelated with $\mathcal{C}$ (Spearman $\rho = -0.13$, $p = 0.27$) while

Table 6: **Sign-forecast agreement by model (IOI fine-tuning).** Amplify-sign agreement per model, listed in order of size. The 2.8–7B and instruction-tuned models use 3 seeds.

Family	Models	Amplify-sign agreement
GPT-2	small, medium, large, xl	1.00, 0.90, 0.97, 0.97
GPT-Neo	125M	0.83
Pythia	410M, 1B, 1.4B, 2.8B, 6.9B	0.77, 1.00, 0.83, 0.80, 1.00
Qwen2.5	0.5B, 1.5B, 3B, 7B	0.60, 0.97, 0.97, 0.77
Qwen2.5-Instruct	3B, 7B	0.90, 0.40

declining with extreme realized gain ($\rho = -0.58$, $p = 5 \times 10^{-8}$): the dissociation is not a step-size artifact, and sign collapse tracks the over-drive boundary, not tangent deviation. An independent kernel-evolution measurement supports the same reading: comparing the base and trained eNTK on the task batch, high-$\mathcal{C}$ runs leave the lazy regime (kernel CKA 0.33–0.49, kernel trace growing 9–$18\times$, versus CKA 0.90 for lazy SFT) while their sign forecasts hold at 0.90–1.00 and calibration error rises with the drift. The one sign collapse (an over-driven negative-gain run, amp-sign 0.00) shows milder kernel drift (CKA 0.71) than the surviving DPO run, so collapse is set by the gain boundary, not by kernel evolution.

The learning-rate sweeps put both tiers on the same 36 runs. Across the 30 productive runs of the two sweeps the structural tier has no variance to explain (amplify-sign is 1.00 on every one), while the first-order calibration error spans 0.0013–0.34 (a factor of 257) and the localization rank correlation drifts only mildly ($0.68 \to 0.62$ and $0.65 \to 0.55$ on the two sweeps): the dissociation is visible without leaving a single experimental design. Decomposing the error growth by cell (sweep $\times$ rate, 3 seeds each) separates a systematic from an erratic component: the cell-mean calibration error rises with the cell-mean $\mathcal{C}$ (Spearman 0.77) and so does the within-cell seed spread (0.56), but the coefficient of variation does not (-0.35), so the first-order prediction becomes systematically mis-scaled rather than seed-unstable, and $\rho(\mathcal{C}, \mathrm{CE})$ reproduces at 0.79 and 0.83 on the two sweeps. Within a learning-rate sweep, however, $\mathcal{C}$ and the realized gain are collinear on productive runs (Spearman 0.74, both driven by the rate), so the sweep alone cannot attribute the metric-tier degradation to tangent deviation rather than to update size (partial $\rho(\mathcal{C}, \mathrm{CE} \mid \mathrm{gain}) = 0.14$ there). The design does establish the contrast between the tiers on identical runs, which is what the dissociation claims. The independence of the structural tier from $\mathcal{C}$ is established separately on the 87-run pool, where the two are decorrelated by construction.

The Gini of base |DLA|, used as the crispness predictor, is a head-count-invariant statistic chosen after the pre-specified top-5 share proved confounded, so it could be a post-hoc choice that tracks architecture size. The controls indicate otherwise. Partialling out, one at a time, total head count, layer count, heads per layer, d_{model}, vocabulary size (a tokenizer proxy), and a parameter-scale proxy leaves the correlation at 0.62–0.69 (permutation $p \leq 0.025$ in every case), and no confound has predictive power of its own (each correlates with amplify-sign between -0.29 and $+0.08$, none significant). It is stable under resampling: leave-one-model-out gives $\rho \in [0.56, 0.76]$ and leave-one-family-out 0.58–0.67, sign-stable throughout. It is not merely a sign statistic: Gini also predicts the localization tier (top-10 overlap, $\rho = 0.59$, $p = 0.034$) while being uncorrelated with the realized task gain (-0.19), so it is not tracking how much the task moved. However, among the concentration statistics we can compute, only the Gini reaches significance (participation-ratio variants and the top-5 share give -0.21 to $+0.39$, none significant), so the choice of inequality measure matters and the paper labels it exploratory. Its one prospective test is the sealed battery, where the cell with the lowest sealed Gini was also the cell with the lowest realized amplify-sign, as pre-declared.

Stronger nulls for the sign forecast

A fair coin is not the right chance level for raw amp-sign when a circuit's predicted signs are unbalanced: if k of the N sealed heads are predicted positive, a sign-blind rule can score well above 0.5. We therefore evaluate three stronger nulls on the frozen runs, using only base-model quantities; compositions k are recomputed on the exact frozen eval cells, per seed.

Majority-sign baseline. The best constant predictor assigns every head the circuit's majority sign and scores $\max(k, N-k)/N$: 0.60 on the canonical IOI circuit (k=3 of 5), 0.60–0.90 across the ten-model sweep and the control tasks. The forecast beats it with margin exactly where the circuit

is sign-mixed (canonical IOI 1.00 vs 0.60, GPT-2 top-10 1.00 vs 0.70, Successor 0.90 vs 0.60, Qwen2.5-0.5B 0.90 vs 0.60) and in both sign-flip regimes below. On sign-pure circuits raw amp-sign has little or no headroom (GPT-Neo and Pythia-1.4b tie the baseline), and Pythia-410m (0.70 vs 0.80) and Greater-Than (0.80 vs 0.90) fall below it. In those cells localization is carried by the overlap, ρ, and sealed-ablation tiers, not by raw amp-sign. Amp-sign also does not track sign purity across cells, which is the signature a covert majority rule would leave.

Composition-preserving permutation test. Permuting the sealed sign labels across the N heads while preserving the predicted composition k makes the agreement count hypergeometric, with chance level $\big(km + (N-k)(N-m)\big)/N^2 \in [0.52, 0.82]$ across our cells (m = realized positive count), not 0.5. We compute exact per-seed p-values, conservatively maximized over the realized compositions consistent with the observed agreement when per-head realized signs were not logged (exact when amp-sign $= 1$), and summarize each cell by its median seed: GPT-2 0.0083, gpt2-large/gpt2-xl/Qwen2.5-1.5B 0.022, Qwen2.5-0.5B 0.033, Pythia-1b 0.10, gpt2-medium 0.20; GPT-Neo, Pythia-410m, and Pythia-1.4b individually cannot reject. Combining the ten independent base models by Fisher's method rejects the permutation null jointly at $\chi^2_{20} = 47.0$, $p = 5.8 \times 10^{-4}$ (the task, reversal, and divergence cells reuse the GPT-2 base and are therefore not pooled).

Balanced-sign agreement. Averaging per-class agreement over predicted-positive and predicted-negative heads restores 0.5 as the exact chance level for any constant or majority rule. The confirmatory cell scores 1.00 (Name Movers $24/24$ up, Negative Name Movers $16/16$ down, 8 seeds, where the minority class is the one in which every constant baseline scores 0), the top-10 heads of the seed-0 run score 1.00 ($7/7$ and $3/3$), and the sealed reversal is bounded below by 0.83.

The same nulls applied to the held-out duality check (H4). The duality battery scores 96 held-out capabilities per seed, each "prefer token A_j over B_j" at the final position of a held-out prompt, with A_j, B_j drawn i.i.d. from the vocabulary: unlike a circuit's sealed signs, the battery is symmetric by construction, and we confirm it is (base-margin positive fraction 0.503, range 0.438–0.562 over the 8 seeds), so 0.5 is the right chance level here. Logging the per-probe signs and applying the three nulls above gives: raw agreement 0.88 against a majority-sign baseline of 0.56 and a composition-matched permutation chance of 0.51; balanced (per-class) agreement 0.88, i.e. indistinguishable from the raw value, so the result is not a majority artifact; and an exact permutation test rejects on every seed ($p \le 5 \times 10^{-5}$). The disagreements are where the noise floor predicts they should be: agreement is 0.95 on probes whose realized change is above the median magnitude and 0.82 below it, so the residual errors are small-movement probes rather than a systematic class of mis-signed capabilities.

Amplify-everything is a rival hypothesis, not a null. The baseline "training amplifies every head along its base sign" coincides with our forecast on-circuit whenever the training gain is positive, so the regimes where the hypotheses disagree decide it, and it loses in all of them: (i) on the sealed reversal it scores 0.10 ($8/8$ seeds) against the sign-negated forecast's 0.90, at or below its own permutation chance (0.38–0.46); (ii) at the divergence boundary (lr $\ge 2.5 \times 10^{-3}$, gain < 0) it scores 0.00 on the canonical circuit while the flip forecast scores $5/5$ in all four runs (exact per-run permutation $p = 0.10$); (iii) off-circuit it is at chance: over the 134 non-forecast heads of the seed-0 run, head-wise sign agreement is 0.515 (binomial vs. coin, two-sided $p = 0.80$), even though 75.7% of the realized $|\Delta\mathrm{DLA}|$ mass lands on the 10 forecast heads. The $|\Delta\mathrm{DLA}|$-weighted agreement is higher (0.88) because movement concentrates near the circuit, but the rival hypothesis is per-head, so the head-wise rate is the relevant statistic. In these runs fine-tuning amplifies the forecast circuit rather than every head. The frozen placebos agree: shuffled-label seals score 0.56 ($\approx$ the analytic permutation chance) and random-readout seals 0.40. A direct head-to-head sharpens the rival comparison: the amplify-everything rule coincides with the sealed forecast on every productive task-improving cell but issues the opposite forecast on the sealed reversal, where it scores 0.10 versus the forecast's 0.90 on all 8 seeds (paired sign test $p = 0.0078$). Across the 87-run pool it falls below 0.5 on 17 runs versus the forecast's 9, all 9 outside the pre-registered validity domain. The outcome-aware variant $\mathrm{sign}(\mathrm{gain}) \cdot \mathrm{sign}(\mathrm{DLA}_{\mathrm{base}})$ is not an independent rival: it is the C2 sign law itself and requires the realized gain, which is unavailable before training.

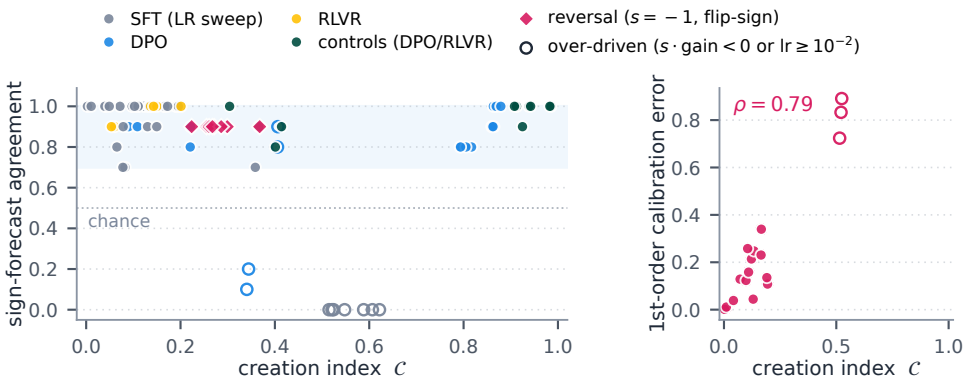

Figure 5: **Structural–metric dissociation over** 87 **runs.** *Left:* top-10 sign-forecast agreement per run versus $\mathcal{C}$ (magenta diamonds: sealed reversal, flip-sign agreement, Corollary 1). Only over-driven runs (open markers: negative training-objective gain or $\text{lr} \geq 10^{-2}$) collapse. *Right:* calibration error rises with $\mathcal{C}$ (ρ=0.79, out-of-sample 0.83).

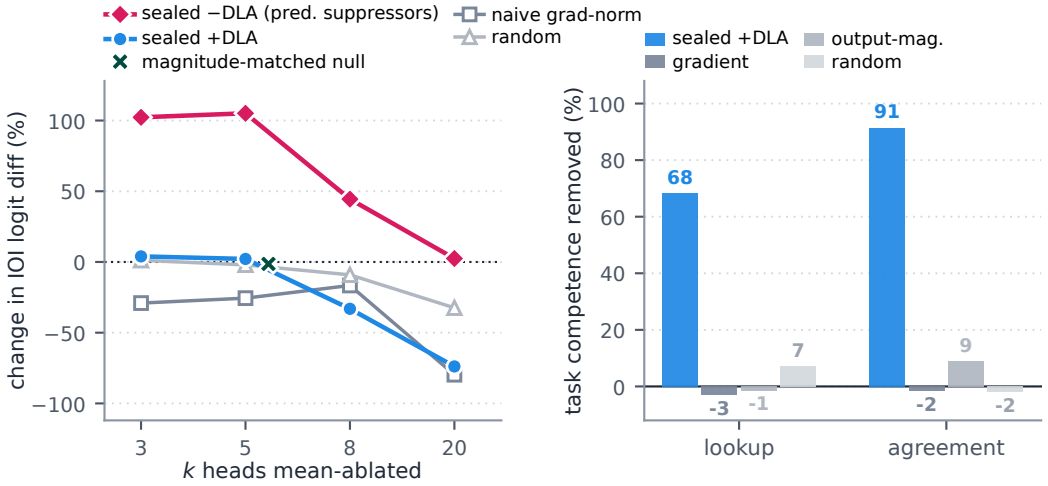

Figure 6: **Causal validation by signed mean-ablation.** *Left:* GPT-2 small IOI. Ablating the sealed −DLA heads improves the task (+105% at k=5). +DLA ablations are masked by self-repair at small k and remove 74% at k=20. *Right:* on the lookup and agreement controls (k=5), the sealed +DLA heads remove 68% and 91% of competence, baselines at most 9%.

C A BASE-TANGENT READING OF THE OUTPUT-SPACE LAWS

On-policy RL defines δ only on outputs the base already samples, so the required v lies in the high-mass cone: $\text{supp}(\delta) \subseteq \text{supp}(\pi_0)$, a base-tangent reading of the support-preservation premise of Wu et al. (2025). Second-order KL gives $\text{KL}(\pi_0 \| \pi_{\theta_0+\Delta\theta}) \approx \frac{1}{2}\Delta\theta^\top F \Delta\theta$ with Fisher $F = J_0^\top H J_0$. The KL-minimal bias of on-policy RL (Shenfeld et al., 2026) is then selection of the min-Fisher-norm $\Delta\theta$, and Cauchy–Schwarz bounds every old capability's forgetting by $\sqrt{\text{Cost}} \cdot \|P_{\text{new}} x_{\text{old}}\|$, consistent with RL's Razor. The KL$\approx \frac{1}{2}\Delta\theta^\top F \Delta\theta$ bridge is textbook, and the base-tangent reading is our contribution.

D PROOFS AND EMPIRICAL CONDITIONS FOR SECTION 3.2

Proof of Lemma 1. Chain rule: $\Delta\text{DLA}_h = a_h(\theta_0 + \Delta\theta) - a_h(\theta_0) = \langle \nabla_\theta a_h(\theta_0), \Delta\theta \rangle + O(\|\Delta\theta\|^2)$. Substituting $\Delta\theta = \eta J_0^\top \delta$ and dropping the second-order term (Assumption 1) gives $\eta\, \kappa_h(0)$. □

Proof of Proposition 1. $a_h(\theta_t)$ is differentiable in t with $\frac{d}{dt}a_h(\theta_t) = \langle \nabla_\theta a_h(\theta_t), \dot{\theta}_t \rangle = \kappa_h(t)$ for gradient flow $\dot{\theta}_t = \sum_i J_t(x_i)^\top \delta_i(t)$, and the fundamental theorem of calculus gives the exact identity. A sign-definite integrand has an integral of that sign, and $\mathrm{sign}(\kappa_h(0))$ is the base-model value. Laziness alone does not imply C1: even with a frozen kernel ($J_t \equiv J_0$, square loss), $\kappa_h(t) = \langle J_0 \nabla a_h, e^{-\Theta t}\delta(0) \rangle$ can cross zero when fast- and slow-mode projections disagree in sign: e.g. eigenvalues $(1, 100)$, $\delta(0) = (1, 1)$ in the eigenbasis, head weights $(-0.005, 1)$ give $\kappa_h(t) = -0.005e^{-t} + e^{-100t}$, positive at $t{=}0$ and crossing at $t \approx 0.05$. What restores C1 is C2: if all circuit-head tangents share one dominant direction, $\nabla a_h \approx g_h w$, then $\kappa_h(t) \approx g_h\, c(t)$ with a common scalar mode $c(t)$ whose sign is the error's readout projection, and C1 holds for every head exactly while C0 does. Conversely C1 tolerates arbitrary kernel drift that preserves the half-space, so it is strictly weaker than laziness+C2. □

Proof of Proposition 2. Write $\delta_i = s\beta_i v_i + \delta_i^\perp$ (C0, $\beta_i \geq 0$). Then $\kappa_h(0) = s\sum_c K_{hc}\bar{\beta} + r_h$, where the residual r_h collects the off-readout error component $\delta^\perp$ and the projections of the error onto non-head tangents (MLP and direct-path components, which the exact decomposition also carries): the row sums of the circuit-head component kernel carry the sign. Under C2, $K \approx \|w\|^2 g g^\top$ with $a_h(\theta_0) = g_h\varphi_0$, $\varphi_0 > 0$, so the row sum is $g_h\, G\|w\|^2$, $G = \sum_c g_c > 0$ (base competence), giving $\mathrm{sign}(\kappa_h(0)) = s \cdot \mathrm{sign}(g_h) = s \cdot \mathrm{sign}(\mathrm{DLA}_{\mathrm{base},h})$ whenever the dominant-mode term exceeds $|r_h|$, a per-head sign margin we measure below rather than bound analytically. □

Empirical conditions (base GPT-2, CPU). Computing $\nabla_\theta a_h$ per circuit head and $\kappa_h(0) = \langle \nabla_\theta a_h, -\nabla_\theta L \rangle$: $\mathrm{sign}(\kappa_h(0)) = \mathrm{sign}(\mathrm{DLA}_{\mathrm{base},h})$ on $12/12$ circuit heads. The 12×12 kernel K has top-eigenvalue share 0.76 (C2 rank-one dominance), and the sign pattern of the dominant eigenvector reproduces the Name Mover/Negative split exactly. The per-head sign margin of Proposition 2 is met on $12/12$ heads: regressing κ on the base attributions, the dominant-mode term exceeds the residual $|r_h|$ for every circuit head (rank-one fit $R^2 = 0.94$), so no head's sign depends on the unmodeled remainder. Beyond sign, the base-model $\kappa_h(0)$ is quantitatively calibrated in the regime where Proposition 1 licenses it: against the realized per-head $\Delta\mathrm{DLA}$ of the sealed SFT run ($\mathcal{C}{=}0.14$), rank correlation 0.93 (Pearson 0.87) across the 12 circuit heads, with signs $12/12$. The first-order tier is thus confirmed directly, not only through its degradation. The objective dependence is also measurable at forecast time: recomputing $\kappa_h(0)$ against the DPO, GRPO-style RLVR, and reversal loss gradients at initialization (no training), $\mathrm{sign}(\kappa_h(0))$ matches $\mathrm{sign}(\mathrm{DLA})$ on $12/12$ heads for all three task-improving objectives (the DPO value on one head, 11.2, the weakest Negative, is numerically zero on the smallest batch and resolves to the predicted sign at larger n), while the reversal gradient flips $11/12$: the sole non-flipping κ (backup head 11.2, whose tangent is dominated by the objective-independent softmax component rather than the IO−S contrast) is exactly the head that fails to flip in the realized reversal (flip rate $9/10$ in every seed, per-head trajectory measurement below). The objective enters the forecast only through the sign of its error-readout projection. C1 itself is measured along the realized trajectories (seed-0 CPU re-runs of the sealed training configurations, checkpoints every 5 steps, with per-run $\mathcal{C}$ values differ slightly from the multi-seed headline means $0.14/0.28/0.87$): $\kappa_h(t)$ retains its initial sign on $12/12$ heads at every checkpoint for SFT ($\mathcal{C}{=}0.19$, seed 0) and, constant throughout, matching the negated forecast on $11/12$ heads under the $s{=}-1$ reversal ($\mathcal{C}{=}0.26$, the non-flipper being 11.2, as forecast at initialization). In the elicitation regime C1 is observed, not assumed. In the creation-heavy DPO run ($\mathcal{C}{=}0.86$) pointwise sign-constancy fails ($0/12$): $\kappa_h(t)$ flips coherently across heads as the shared C2 factor oscillates with the saturating preference margin, yet the realized $\Delta\mathrm{DLA}$ (the exact transport integral of Proposition 1) still lands on the sealed sign on $11/12$ heads (the exception is again 11.2, while a coarse 5-step discretization of $\int \kappa\, dt$ under-samples the oscillation and gives $7/12$). C1 thus separates the regimes pointwise, while the path integral of Proposition 1 is what the forecast actually needs. The same structure replicates on a second task: on lookup, where even SFT is creation-heavy ($\mathcal{C}{=}0.88$, and DPO 0.98), pointwise sign-constancy fails at most checkpoints yet the integrated transport $\int \kappa_h\, dt$ and the realized final $\Delta\mathrm{DLA}$ both match the sealed signs on $12/12$ heads in both runs, so the path-integral account is not specific to IOI. The reversal fine-tune (Section 4.1) instantiates $s{=}-1$: flip-sign 0.90 per seed at $\mathcal{C} = 0.28$, localization set unchanged, while the amplify-prior score drops to 0.10.

The survival law. The trajectory measurements above motivate a stronger condition than C1, under which the sign forecast survives without pointwise sign-constancy.

Proposition 6 (signed transport under persistent factorization)**.** *Let θ_t, $t \in [0,T]$, be any absolutely continuous training path and $\kappa_h(t) = \langle \nabla_\theta a_h(\theta_t), \dot\theta_t \rangle$ the transport integrand of Proposition 1, so that $\Delta\mathrm{DLA}_h = \int_0^T \kappa_h(t)\,dt$ exactly. Suppose the integrand factorizes with fixed loadings,*

$$\kappa_h(t) = g_h\, c(t) + r_h(t) \qquad (\textbf{\textit{C2-persistence}}),$$

and head h satisfies the margin $\int_0^T |r_h|\,dt < |g_h|\,\big|\int_0^T c\,dt\big|$. Then

$$\mathrm{sign}(\Delta\mathrm{DLA}_h) = \mathrm{sign}(g_h) \cdot \mathrm{sign}\Big(\int_0^T c\,dt\Big),$$

with no pointwise sign condition: c may cross zero arbitrarily often, so C1 may fail on every head at every instant. Unconditionally, $\big|\Delta\mathrm{DLA}_h - g_h \int c\big| \le \int |r_h|$, so as $\int c \to 0$ the transport of every head collapses to its remainder scale simultaneously.

Proof. $\Delta\mathrm{DLA}_h = g_h \int c + \int r_h$ and $|\int r_h| \le \int |r_h| < |g_h \int c|$, so $\Delta\mathrm{DLA}_h$ lies strictly on the side of $g_h \int c$. □

Lemma 2 (integrated gain identification (C0-integrated))**.** *Let the task margin decompose exactly as $m(\theta) = \sum_h a_h(\theta) + a_\perp(\theta)$ ($a_\perp$: non-circuit terms) and let $G = m(\theta_T) - m(\theta_0)$ be the realized task gain. If (i) $\sum_h g_h > 0$ (net-positive circuit loading; base competence) and (ii) $\sum_h \int |r_h| + |\Delta a_\perp| < \big(\sum_h g_h\big)\big|\int c\big|$ (circuit dominance), then $\mathrm{sign}\big(\int c\big) = \mathrm{sign}(G)$, and hence, under the margins of Proposition 6,*

$$\mathrm{sign}(\Delta\mathrm{DLA}_h) = \mathrm{sign}(g_h) \cdot \mathrm{sign}(G) \quad \textit{for every head:}$$

per-head transport signs from fixed loadings and one realized scalar, the final gain.

Proof. $G = \sum_h \Delta\mathrm{DLA}_h + \Delta a_\perp = \big(\sum_h g_h\big)\int c + \sum_h \int r_h + \Delta a_\perp$, and (ii) makes the last two terms a strict minority. □

Corollary 3 (collapse exactly at the gain sign-flip)**.** *By the identity in Lemma 2, $\big|\int c\big| \le \big(|G| + \sum_h \int |r_h| + |\Delta a_\perp|\big) / \sum_h g_h$, so along any family of runs driven through $G \to 0$ with bounded remainders, the predicted transport $g_h \int c$ shrinks on all heads simultaneously and every realized $|\Delta\mathrm{DLA}_h|$ falls below $\int |r_h| + (|g_h| / \sum_c g_c)\big(\sum_c \int |r_c| + |\Delta a_\perp|\big)$, its own remainder plus a $|$loading$|$-weighted share of the collective one: the forecast does not fail head-by-head but all-at-once, exactly at the gain sign-flip, and on the far side every head's sign flips coherently with G. (Conversely, under condition (ii), $|G| \ge \big(\sum_h g_h\big)\big|\int c\big| - \sum_h \int |r_h| - |\Delta a_\perp| > 0$. When (ii) fails near the flip, the theory predicts near-zero, sign-fragile transport rather than a sign.) In the exact instantiations below $r \equiv 0$, $a_\perp \equiv 0$ and $\Delta\mathrm{DLA}_h = g_h G / \sum_c g_c$: collapse and coherent flip are exact.*

Theorem 2 (exact factorization in tractable models)**.** *C2-persistence holds with $r_h \equiv 0$, for any step size, in the following models.*

*(i) **Rank-one-kernel descent (lazy).** Under Assumption 1 with rank-one base Jacobian $J_0 = uw^\top$ (eNTK $\Theta_0 = \lambda uu^\top$, $\lambda = \|u\|^2\|w\|^2$) and square loss with output error δ_k, every observable a_h has*

$$\kappa_h(k) = g_h\, c_k, \qquad g_h = \langle \nabla_\theta a_h(\theta_0), w \rangle, \qquad c_k = u^\top \delta_k = (1-\eta\lambda)^k\, c_0,$$

i.e. the shared factor is the single decaying error mode, and $\Delta\mathrm{DLA}_h = \eta g_h \sum_k c_k$ exactly, and the realized readout gain is the fixed positive multiple $G = (v^\top u)\|w\|^2 \eta \sum_k c_k$ whenever $v^\top u > 0$ (C0). For $\eta\lambda \in (1,2)$ the mode overshoots: c_k alternates sign every step (C1 fails on every head) while $\sum_k c_k$ keeps the sign of c_0, so signed transport survives. For $\eta\lambda > 2$ the iteration diverges, G flips sign with the parity of the step count, and all heads flip with it, with $\Delta\mathrm{DLA}_h \propto g_h G$ exactly: collapse of every head exactly at $G = 0$.

*(ii) **Deep-linear feature learning (non-lazy).** Let $f(x) = W_2 W_1 x$ with whitened inputs, teacher $T = W_2^* W_1^* + s\,uv^\top$ (background exactly fitted, residual rank-one), and aligned initialization $W_2(0) = W_2^* + \alpha_0 uw^\top$, $W_1(0) = W_1^* + \beta_0 wv^\top$ where the background is silent on the support of the recruited hidden direction w: $W_2^* e_h = 0$ and $e_h^\top W_1^* = 0$ for every $h \in \mathrm{supp}(w)$ (hence $W_2^* w = 0$, $w^\top W_1^* = 0$), and $W_2^{*\top} u = 0$, $W_1^* v = 0$. Then discrete gradient descent of any step size preserves the aligned form exactly, $\alpha_{k+1} = \alpha_k + \eta\beta_k(s - m_k)$, $\beta_{k+1} = \beta_k + \eta\alpha_k(s - m_k)$, $m_k = \alpha_k\beta_k$, and*

for the per-head reads a_h (hidden-unit contributions $a_h = (q^\top W_2 e_h)(e_h^\top W_1 \bar{x})$ to the probe task logit, which decompose it exactly, or per-coordinate output reads $a_h = q_h\, e_h^\top W_2 W_1 \bar{x}$) the transport integrand factorizes exactly,

$$\kappa_h(k) = g_h\, c_k, \qquad c_k = (\alpha_k^2 + \beta_k^2)(s - m_k),$$

with fixed loadings $g_h = (q^\top u)(v^\top \bar{x})\, w_h^2$ (resp. $q_h u_h (v^\top \bar{x})$, of mixed signs), even though the tangent features $\nabla a_h(\theta_k)$ themselves move: the run is not lazy. Moreover the realized changes factorize exactly at any step size, $\Delta a_h = g_h(m_K - m_0)$ and $G = (q^\top u)(v^\top \bar{x})(m_K - m_0)$, so $\operatorname{sign}(\Delta \mathrm{DLA}_h) = \operatorname{sign}(g_h)\operatorname{sign}(G)$ with no smallness anywhere; base attributions obey $a_h(0) = g_h m_0$ (+ background constant for output reads): where the background writes nothing on head h, loadings are proportional to base attributions: C2's alignment clause, derived. Under gradient flow $\dot{m} = c(t)$ (balanced init: $\dot{m} = 2m(s-m)$), so $\int c = m(T) - m(0)$ is the net motion of the learned mode. At large η, m_k overshoots s and c_k oscillates (C1 fails pointwise on every head) while signed transport holds, with all heads collapsing and flipping together exactly at $G = 0$.

Proof sketch. Every algebraic identity below is additionally verified numerically ($35/35$ float64 checks at 10^{-14}–10^{-19}). (i) $\dot{\theta} = \eta J_0^\top \delta = \eta w (u^\top \delta)$, so $\kappa_h = \langle \nabla a_h, w \rangle (u^\top \delta)$. Under Assumption 1 applied componentwise ∇a_h is fixed, $\delta_{k+1} = (I - \eta \Theta_0)\delta_k$ leaves $\delta^\perp := (I - uu^\top/\|u\|^2)\delta$ invariant and multiplies the u-mode by $(1 - \eta\lambda)$, so the sums are geometric. (ii) With the four orthogonality conditions, $W_2 W_1 = W_2^* W_1^* + m\, uv^\top$, hence $E = T - W_2 W_1 = (s - m)uv^\top$ stays rank-one. The GD updates are $-\eta \nabla_{W_2} L = \eta E W_1^\top = \eta\beta(s-m)uw^\top$ and $-\eta \nabla_{W_1} L = \eta W_2^\top E = \eta\alpha(s-m)wv^\top$, so the aligned form is preserved with the stated recursion (the manifold is invariant, exactly, at any η, in exact arithmetic). Direct substitution gives $\kappa_h = \langle \nabla a_h, -\nabla L \rangle = g_h c_k$ for both head reads (for hidden-unit reads the background cross-terms $(s-m)w_h[\beta(q^\top u)(e_h^\top W_1^* \bar{x}) + \alpha(v^\top \bar{x})(q^\top W_2^* e_h)]$ vanish per head because the background is silent on $\operatorname{supp}(w)$, whereas the weaker conditions $W_2^* w = 0$, $w^\top W_1^* = 0$ alone annihilate only their sum over heads), and $a_h(k) = g_h m_k (+\mathrm{const}_h)$ gives the realized factorization, and summing heads gives G. The discrete-time correction to the flow identity is the exact relation $m_{k+1} - m_k = \eta c_k + \eta^2 m_k (s - m_k)^2$. □

Remark 1 (boundary of exactness)**.** *If the recruited mode shares hidden units with the background ($W_2^* w \neq 0$ or $w^\top W_1^* \neq 0$), the initial error is no longer rank-one, the aligned manifold is not invariant, and the tangent matrix $[\kappa_h(k)]$ acquires macroscopic higher singular values (top-1 energy 0.95 in our instance): C2-persistence is an alignment property of the recruited mode, not a universal consequence of gradient training. The measured rank-one energies on the realized trajectories quantify this property (0.890–0.999), and its failure is the mechanism behind over-drive breaking the sign at positive gain (C2 failure). The aligned manifold, while exactly invariant, is also numerically unstable at large η. The over-driven checks project each iterate back onto it to control float64 drift ($\sim 10^{-16}$ per step), and in finite precision (as in real training) over-drive itself pushes trajectories off the aligned manifold. Softmax attention and LayerNorm are outside the scope of the exact statement.*

Measured C2-persistence along the realized trajectories. Stacking the checkpointed $\kappa_h(t)$ of the five seed-0 re-runs (IOI SFT/reversal/DPO; lookup SFT/DPO) into heads×time matrices, the top-singular-vector energy share is $0.999/0.995/0.890$ and $0.993/0.992$ respectively (column-normalized: $0.94/0.95/0.78/0.88/0.95$). The per-checkpoint $|\cos(\kappa(t), g)|$ median varies across runs (0.66 IOI-SFT, 0.89 reversal, 0.95 DPO, $0.88/0.98$ lookup). The shared factor $c(t)$ rotates and crosses zero along the path, so pointwise alignment is not maintained, as Proposition 6 permits (no pointwise sign condition). The endpoint forecast uses the persistent loading g and the net gain, not the pointwise cosine. The low-cosine checkpoints are the remainder-dominated zero-crossings the proposition predicts, and the DPO 0.21 outlier sits at one such crossing. The extracted loadings g (oriented by the base model alone) match $\operatorname{sign}(\mathrm{DLA}_{\mathrm{base}})$ on $12/12$, $11/12$, $11/12$, $12/12$, $12/12$, and the sole mismatches are head 11.2, whose DPO loading is the smallest of all twelve. Gain identification is $5/5$ on a held-out eval batch of the same runs (loadings and factor come from the kappa batch and realized $\Delta\mathrm{DLA}$ from a disjoint eval batch, so the split is across batches, not runs). The slope of the realized held-out $\Delta\mathrm{DLA}$ on g carries the sign of the realized training gain on every run (R^2 through the origin 0.76–0.90). The derived endpoint predictor $\operatorname{sign}(g_h) \cdot \operatorname{sign}(G)$ then reproduces the realized $\Delta\mathrm{DLA}$ sign on $60/60$ head–run pairs, turning the 11.2 anomaly of the sealed forecast into the persistent-factor prediction ($g_{11.2} = +0.019$, realized $+0.209$ under DPO;

$g_{11.2} = +0.085$, gain -11, realized -1.51 under reversal). The strict per-head margin of Proposition 6 is met, in the endpoint decomposition $\Delta\mathrm{DLA}_h = g_h\hat{G} + \mathrm{resid}_h$, on $8/11/11/10/10$ of 12 heads per run, and the failures land on the predicted sign via reinforcing remainders (e.g. head 7.9 under DPO, ratio 1.30). The checkpoint-grid quadrature of $\int c$ is sign-correct on $4/5$ runs. The exception is the DPO oscillation, aliased by the 5-step grid as noted above for $\int \kappa$.

E PRE-REGISTRATION AND INTEGRITY

This appendix separates the hypotheses fixed before fine-tuning from the analyses added later.

Sealing protocol. The confirmatory hypotheses (H1b, H2, H3, H4, causal self-repair) and their acceptance thresholds were fixed for GPT-2 small, IOI, and SFT with the NLL loss before any fine-tuning run. IOI was chosen because it was a priori the most thoroughly documented GPT-2 circuit, not because it scores best: the Greater-Than and Successor circuits attain higher rank correlations (0.85 and 0.77, against 0.65 for IOI). The other circuits, models, and objectives are exploratory replications of the same prediction form. The per-head predictions are computed on CPU from the base model before any fine-tuning run.

Confirmatory hypotheses. **H1b:** top-5 ΔDLA heads include ≥ 2 of the Name Movers $\{9.9, 9.6, 10.0\}$ and Spearman $\rho(|\text{base DLA}|, |\Delta\text{DLA}|) \geq 0.60$. **H2:** $\mathrm{sign}(\Delta\text{DLA}) = \mathrm{sign}(\text{base DLA})$ ($+$ for Name Movers, $-$ for Negative Name Movers). **H3:** in-support target reachable $\gg$ out-of-context-name target at matched LR. **H4:** $\mathrm{sign}(\tilde{u}^\top \tilde{\Theta}_{\text{old,new}}\alpha)$ predicts held-out capability-change sign (≥ 0.75). **Causal:** ablating $-$DLA heads raises IOI, ablating the $+$DLA set lowers it.

Deviations from pre-registration. (i) The localization estimator was changed, before any fine-tuning run, from the raw parameter-gradient $\|P_c J_0^\top \delta\|$ to the direct logit attribution (DLA), after CPU analysis on the base model showed that the raw gradient does not localize. The raw gradient is reported as a baseline that fails to localize. (ii) The tautological metric $\rho(a, |\Delta W_O|)$ was demoted in favor of $\rho(|\text{DLA}|, |\Delta\text{DLA}|)$, also pre-run. (iii) The pre-registered optimizer settings were adjusted for stability before the first run (learning rate 10^{-3}, 50 steps, and 512 prompts became 3×10^{-4}, 40 steps, and 256–96 prompts), without changing the acceptance thresholds. (iv) The pre-registered H4 estimator was the base cross-task Fisher-eNTK sign $\mathrm{sign}(\tilde{u}^\top \tilde{\Theta}_{\text{old,new}}\,\alpha)$, whereas the executed implementation evaluates the base tangent against the realized update direction ($J_0\Delta\theta$, one JVP). The reported 0.88 is therefore a base-tangent consistency check computable only after the run, like $\mathcal{C}$, with the pre-registered $\geq 75\%$ acceptance threshold unchanged. This substitution was present in the implementation from the first run. The implementation also excludes probes whose realized change is sign-undefined at noise level ($|\Delta| \leq 0.2\,\mathrm{std}$, retaining 76–87 of 96 probes per seed), a floor not stated in the pre-registered criterion. No thresholds were changed after seeing results, and no seeds were dropped. The sealed reversal forecast is the deterministic sign negation of the original forecast (checked on all 144 heads), so it introduces no new degrees of freedom.

Exploratory analyses. The second and third circuits (Greater-Than, Successor); cross-model generalization (gpt2-{medium,large,xl}, pythia-410m, the 2.8–7B and instruction-tuned spot-checks); objective invariance (DPO, RLVR, GRPO, multi-token); the creation index $\mathcal{C}$ and its regime classification; and the dissociation between localization and laziness (Section 4.2). These were run after the confirmatory hypotheses were fixed and share the same prediction form, but are reported as exploratory.

Reported experiments. The experiments reported in this paper are the IOI confirmation (8 seeds), the Greater-Than and Successor replications, the ten-model sweep, the causal ablation ($k \in \{3, 5, 8, 20\}$, the self-repair series), reachability (H3) and duality (H4), the dense creation sweep behind the dissociation (40 runs, GPT-2 and Pythia-1b), the crispness test (10 models), the independent-localizer controls (AtP, scale-gauged gradient, output-based), the scale and instruction-tuned spot-checks, the one-token and $K{=}4$ generative GRPO runs, and the seed-bootstrap CIs (Table 7). All CPU theory checks pass, and all experiments use a single GPU. The localization reference (naive gradient-norm head rank, 93 ± 17 vs. DLA 2.1 ± 1.8, 5 seeds) is the reference for all ρ cells.

Table 7: **Seed-bootstrap confidence intervals.** 95% CIs for the primary numbers (10k resamples of per-seed values). Elicitation (SFT and RLVR) and creation (DPO), and in-support and out-of-support reachability, are separated by non-overlapping intervals.

Quantity	Point	95% CI
Localization $\rho(\|\text{DLA}\|, \|\Delta\text{DLA}\|)$ (8 seeds)	0.65	$[0.63, 0.68]$
IOI gain (logits, 8 seeds)	$+4.80$	$[4.70, 4.89]$
H4 held-out duality check (8 seeds)	0.88	$[0.86, 0.91]$
H3 reachability $\mathcal{C}$ in-support	0.03	$[0.03, 0.04]$
H3 reachability $\mathcal{C}$ out-of-support	0.76	$[0.73, 0.80]$
Creation $\mathcal{C}$: SFT	0.14	$[0.11, 0.19]$
Creation $\mathcal{C}$: DPO	0.87	$[0.86, 0.88]$
Creation $\mathcal{C}$: RLVR	0.16	$[0.14, 0.20]$
amp-sign (GPT-2 IOI, all objectives)	1.00	$[1.00, 1.00]$

Prospective sealed replication battery. After the method and all analyses were frozen, we pre-registered and ran a four-cell replication battery under the same protocol. The gates, per-cell base forecasts, and a binding interpretation rule were fixed before any run, and no seed augmentation or re-runs were permitted. The cells were chosen a priori by base-only diagnostics. Three high-confidence cells passed a pre-declared applicability gate (base attribution Gini ≥ 0.75) on distributions sharing no strings with any prior run: a fresh-vocabulary lookup task on GPT-2 small and medium, and an object-relative agreement family, which replaced a first template family that failed the gate under a pre-declared fallback. The fourth cell was pre-declared weak (Qwen2.5-0.5B, the weakest base of the ten-model sweep, on a task–model pair absent from all earlier runs). The battery endpoint is prediction of the validity domain, not all-pass, and all three endpoints held. The three high-confidence cells pass the primary gates (seed-mean amplify-sign 0.89, 0.80, and 0.90, overlap 6.4–7.0/10, ρ permutation $p<0.05$, 8 seeds each). The pre-declared weak cell is the minimum (0.775, below the 0.8 gate and scored FAIL by the preregistered rule), as forecast by its lowest base Gini (0.729 against 0.771–0.867 for the other cells). In the weak cell, localization holds in $8/8$ seeds while the sign gate fails, the same pattern of structural robustness seen elsewhere. The sign forecast again holds under creation-heavy updates ($\mathcal{C} = 0.91$–0.95 on the lookup cells). Three caveats apply. The causal secondary (ablation $\geq 40\%$ and $\geq 3\times$ the best baseline) met its 40% arm (53% and 90% removed) but not the $3\times$ arm ($2.2\times$ and $2.7\times$ the best baseline). The forecast signs of the medium-model cell are all positive, so its sign gate does not by itself separate the forecast from a constant rule, and its evidential weight lies in localization and model generalization. The own-scale secondary of the weak cell passes (0.81), which places it on the scale-convention-sensitive row of the preregistered interpretation rule (realized residual-RMS ratio 2.7).