THE ORDER EFFECT AND THE ORDER IMPRINT ARE NOT THE SAME THING

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-training a language model with two capability operators in different orders produces measurably different models, and a growing line of work uses Lie-bracket surrogates to *predict* which order scores better. We show that this validation target is largely the wrong object. We compose two on-policy distillation operators (math and code teachers) from a shared start over six seeds, and find that one intervention—resetting optimizer state at the phase boundary—pulls apart two quantities the literature treats as one. It collapses the *benchmark* order effect: ΔJ_M falls from -0.0107 to -0.0038 ± 0.0027, inside our preregistered indistinguishable-from-zero band, and uniform weight averaging of the two single-operator models matches the better order in any case. The same intervention leaves the *parameter* order imprint $u = \theta_{MC} - \theta_{CM}$ within 4% of unchanged on every seed. An audit of that imprint shows that its "recency" share depends on a free parameter no prior work controls, namely where operator displacements are measured: 2–3% of $\|u\|^2$ at initialization, 12–15% at the phase boundary. The rest survives every null we can build: it replicates across seeds ($\bar{\varphi} = 0.44$), an SFT pair on the same data leaves a near-orthogonal imprint, random directions of equal norm are inert, and its geometry replicates on a 4B student. Benchmarks thus register mostly optimizer bookkeeping and miss the part of the order effect that replicates. We release the audit protocol as a measurement standard for composition claims. Code is available at `https://anonymous.4open.science/r/pdc-2027`.

1 INTRODUCTION

Sequential post-training is the default way capabilities are added to language models: a base model is instruction-tuned (Ouyang et al., 2022), then aligned (Rafailov et al., 2023), then specialized (Lambert et al., 2024), each stage initialized from the last. When two stages U_A and U_B do not commute, the order of composition leaves a mark: $U_B(U_A(\pi_0)) \neq U_A(U_B(\pi_0))$. Recent work treats this non-commutativity as a *scheduling* problem, using first-order Lie-bracket surrogates to predict which order scores better (Rukhovich et al., 2024; Sweeney, 2026a; Piontkovskaia & Nikolenko, 2026). All of it is validated against benchmark rankings.

This paper reports that benchmark rankings and what actually differs between two orders come apart under a single controlled intervention, and that the mundane explanations one would want to rule out first have gone unmeasured. Our starting observation is a dissociation. Reset the optimizer state at the phase boundary—change nothing else—and the benchmark order effect drops below detectability, while the parameter-space difference between the two composition endpoints does not move. The two quantities respond to the same intervention in opposite ways, so they are not two views of one phenomenon.

That matters because on-policy composition has an order channel that fixed-dataset composition structurally lacks. When the operator is on-policy—the student samples its own rollouts and is pulled toward a teacher on them—the order of composition changes *which states the policy visits*, and this occupancy term is absent by construction when the data distribution does not depend on θ (the leading bracket term of Piontkovskaia & Nikolenko (2026) vanishes identically in that case). An order effect in this regime is therefore the interesting case, and it is the case we build a testbed for.

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.

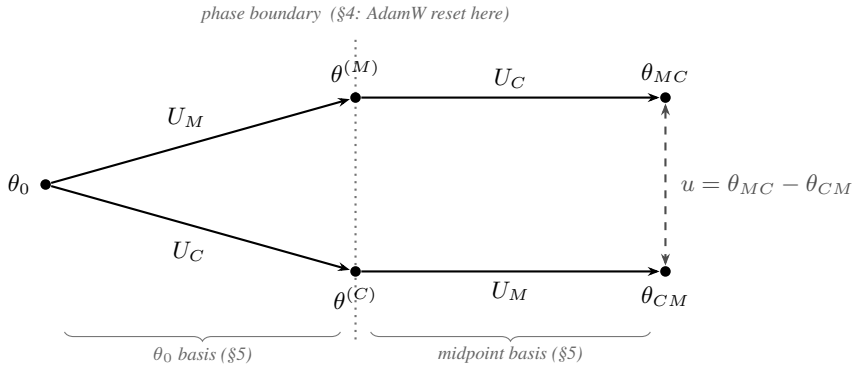

Figure 1: **The composition matrix.** Both orders start from the shared $\pi_0 = \theta_0$; the phase-boundary midpoints $\theta^{(M)}, \theta^{(C)}$ are the half-budget single-operator models. The order imprint u is the difference of the two endpoints. The two braces mark the free parameter of §5: the same operator's displacement can be measured from θ_0 or from the midpoint where it actually acts, and the two give different answers. Not shown: full-budget single-operator endpoints θ_M, θ_C and their merge (§3), and the optimizer-reset arms, which re-run the second phase only.

We compose two on-policy distillation (OPD) operators—toward a math teacher and a code teacher—in both orders from a shared warm-started 1.7B student, across six seeds, with preregistered criteria, matched-seed nulls, and saved completions for every evaluation (§2). The *order imprint* is the parameter-space difference $u = \theta_{MC} - \theta_{CM}$. The paper then has two acts.

Act I: nothing durable in benchmark space (§3). The math-accuracy order effect is modest to begin with ($\Delta J_M = -0.0146 \pm 0.0053$ over six seeds, 2.8 se, sign-consistent in 5/6, with one outlier seed). Resetting optimizer state shrinks it to about a third, leaving a residual of -0.0038 ± 0.0027 that passes the preregistered $|\Delta J_M| \leq 0.006$ line for "indistinguishable from zero": most of what a benchmark sees as order is Adam moment buffers crossing the stage boundary. What remains is not worth having either: uniform weight averaging of the two single-operator endpoints is statistically indistinguishable from the better order and from joint training. Two prescriptions follow directly, and they are the paper's cheapest deliverables.

The pivot (§4). The same reset leaves $\|u\|$ within 4% of unchanged on every seed ($\rho_\theta = 1.007$), and the reset imprint points in the same direction, partially rotated ($\cos = 0.53$–0.57). Whatever the imprint is, it is not optimizer bookkeeping, and it is not what the benchmark was tracking.

Act II: auditing the imprint (§5–§6). The deflationary account of u is recency: each endpoint is dominated by whichever operator ran last. Whether the data support this depends on a free parameter that, to our knowledge, no prior work controls: *where* the operator displacements defining recency are measured. Measured at initialization, they explain 2–3% of $\|u\|^2$; measured at the phase-boundary midpoint, where the second operator actually acts, 12–15% (29–33% jointly). A paper that asserts or rejects recency from one basis is reporting the basis, not the model. What no basis reaches is a direction we then try to kill five ways and cannot: it replicates across seeds, it is operator-specific rather than data-specific, it is functionally a narrow ridge, it sits in Fisher-flat directions, and it replicates at 4B.

The occupancy share itself is the quantity we most wanted and did not get. We report the estimator work in full (§7), including a single-point answer that looked decisive until an ε-scan exposed it as bf16 rounding, because the variance budget we measured is what the next attempt needs.

Our contribution is an *audit standard*, not a training method: a dissociation of benchmark order effect from parameter order imprint, a decomposition of the imprint into basis-dependent recency and a replicable remainder, and the preregistration, null, and gate machinery that makes each attribution stick (checklist in Table 2).

Table 1: Operator-suitability gate history. Each row is a preregistered go/no-go that failed before its arm produced analysis data.

Arm	Gate	Measured	Disposition
1.5B teachers	teacher improves own domain	$\Delta J_C < 0$ (code)	swap to 7B teachers
0.6B student	domain measurable at base	code 0.011 (< 0.03)	line abandoned
4B student	own-domain gain $\geq +0.05$	$+0.027$ / -0.006	scope $\rightarrow \theta$-geometry
		(1.7B arm: $+0.020$, also $<$ gate)	gate flagged miscalibrated
gen$\times$math pair	$\Delta J_G \geq +0.0449$	-0.0018; trunc. $0.32 \rightarrow 1.00$	arm terminated

2 SETUP: OPERATORS, MATRIX, AND APPLICABILITY CONDITIONS

Student and starting point. The student is Qwen3-1.7B-Base (Yang et al., 2025). To give both operators a common format-capable origin, we first build π_0 by short off-policy SFT toward Qwen2.5-1.5B-Instruct on UltraChat prompts (Ding et al., 2023); π_0 establishes the chat template only and is shared by every run.

Operators. A capability operator U_A is K steps of on-policy distillation (Hinton et al., 2015; Agarwal et al., 2024; Gu et al., 2024): the student samples rollouts and minimizes full-vocabulary KL against the teacher's logits on its own samples (Qwen2.5 and Qwen3 share a tokenizer, so distillation is exact in logit space). U_M distills toward Qwen2.5-Math-7B-Instruct (Yang et al., 2024), U_C toward Qwen2.5-Coder-7B-Instruct (Hui et al., 2024).

Composition matrix. From π_0, with total budget T split equally (T=400 steps, 200 per phase, is the primary budget; T=500, 600 secondary): θ_{MC} (math then code), θ_{CM} (code then math), the phase-boundary midpoints $\theta^{(M)}, \theta^{(C)}$, single-operator full-budget runs, joint mixed-batch training, and merging baselines (Figure 1). Six seeds s_0–s_5 vary rollout sampling and data order; seed pairs give matched nulls (Bouthillier et al., 2021). The order imprint is $u_s = \theta^{(s)}_{MC} - \theta^{(s)}_{CM}$. One seed, s_5, is an outlier on every statistic we track—$\|u\|$, ΔJ_M, and stopping behavior alike; we report it throughout rather than trimming it, and mark where it moves an aggregate.

Metrics. J_M: GSM8K (1319; Cobbe et al., 2021) + MATH-500 (Hendrycks et al., 2021; Lightman et al., 2023), symbolic equivalence of boxed answers; J_C: HumanEval+ (164) + MBPP+ (378) (Chen et al., 2021; Austin et al., 2021; Liu et al., 2023), execution-based; J_{avg} their mean. Because answer-format compliance depends strongly on the last operator, all headline J values use a modality-agnostic rescoring pass (extract-then-verify irrespective of format), validated by a shuffled-gold negative control that scores 0.0000 on every condition. The two scorings diverge only on the J_M of C_{half}; Appendix B.4 quantifies the gap.

Applicability: four gate failures in five configurations. A composition audit presupposes that each operator is a policy improvement in its own right. We enforced this as preregistered go/no-go gates, and most candidate configurations did not pass (Table 1): the math$\times$code OPD pair on a 1.7B student is what survived, and it bounds every quantitative claim below. One incident cuts against us and we state it plainly. The 4B arm failed the own-domain-gain gate as written, and the post-hoc audit found the *gate* miscalibrated: it was set on an early pilot and never passed by our flagship 1.7B arm either. Per preregistration, the 4B matrix was scope-reduced to θ-geometry claims (§6v) with a ledger entry. The plain reading is that own-domain benchmark gains at these budgets are small everywhere, which is itself part of what §3 reports; the imprint measurements do not depend on that gate.

Measurement discipline. Every claim in this paper was preregistered (Nosek et al., 2018) with frozen thresholds before the deciding computation ran; every free parameter (thresholds, scoring heuristic, basis choice) is scanned over at least two values; every evaluation saves raw completions, which let us repair three scoring bugs offline without retraining; truncation rates and format compliance are logged as first-order quantities alongside accuracy. Deviations are recorded in a change ledger. Several early conclusions did not survive this discipline, and we report them as results (§7).

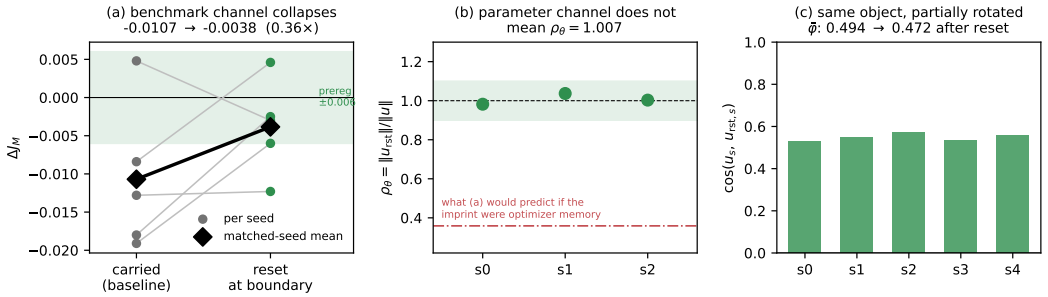

Figure 2: **One intervention, two opposite answers.** (a) Resetting optimizer state at the phase boundary drops ΔJ_M into the preregistered indistinguishable-from-zero band (shaded); grey lines are the five matched seeds. (b) The same intervention leaves the parameter imprint's norm within 4% of unchanged on every seed; the dash-dotted line marks where ρ_θ would sit if the imprint were the same object the benchmark was tracking. (c) It is also the same *direction*, partially rotated, and the reset imprint keeps its cross-seed consistency ($\bar{\varphi} = 0.472$ vs. 0.494 at baseline).

3 Act I: The benchmark order effect is fragile and buys nothing

The raw order effect on math accuracy at $T{=}400$ is $\Delta J_M = J_M(\theta_{MC}) - J_M(\theta_{CM}) = -0.0146 \pm 0.0053$ (mean $\pm$ se over six seeds): about 2.8 standard errors, sign-consistent in 5/6 seeds, with large seed heterogeneity (per-seed values range from $+0.0048$ to -0.0342, the latter the outlier s_5). On the combined metric J_{avg} it is -0.0052 ± 0.0022 (2.4 se, 5/6). The phenomenon a practitioner would observe is therefore real but modest, and concentrated in the domain metric.

Most of it is optimizer state. We rerun the second phase of both orders with AdamW state (moments, step counters; Loshchilov & Hutter, 2019) reset at the phase boundary, changing nothing else (five seeds, preregistered). On the five seeds where both arms exist, the effect shrinks from -0.0107 to -0.0038 ($0.36\times$; on the three seeds the criterion was frozen on, $-0.0166 \to -0.0069$, $0.42\times$). The reset residual -0.0038 ± 0.0027 clears the preregistered $|\Delta J_M| \le 0.006$ line for "the accuracy channel is optimizer memory" and sits 1.4 se from zero. We state the limit of this result as carefully as the result: the *paired reduction* is $+0.0069 \pm 0.0045$, only 1.5 se, and it is not uniform—it vanishes at s_2 and reverses at s_3. What five seeds support is "after the reset there is no detectable order effect," not "the reduction is individually significant." This extends the single-task shuffle-order mechanism of Sweeney (2026b) to cross-operator composition, where it had not been measured.

What is left, averaging gets for free. If order carried usable information, the better composition should beat order-free baselines. It does not. Uniform weight averaging $\theta_{\lambda=0.5} = \frac{1}{2}(\theta_M + \theta_C)$ of the two full-budget single-operator endpoints is statistically indistinguishable from the better order and from joint mixed-batch training on J_{avg}: paired differences over six seeds are $+0.0053 \pm 0.0034$ (vs. the better order, 5/6 positive) and $+0.0020 \pm 0.0042$ (vs. joint, 4/6; seed sd 0.0084 and 0.0102), both under 2 se. We explicitly do *not* claim averaging is better, only not worse. The comparison is calibrated by its own controls: the same pipeline resolves $\lambda = 1.0$ task-vector summation (Ilharco et al., 2023) as reliably worse than $\lambda = 0.5$ (-0.0129 ± 0.0016, 7.9 se, 6/6) and worse than the better order (-0.0077 ± 0.0027, 6/6), with severe generation degeneration—so the null at $\lambda = 0.5$ is not lack of power against coefficient-scale effects. These conclusions were declared on three seeds and confirmed on three fresh seeds.

Two prescriptions. If order sensitivity at the benchmark level is the concern, reset optimizer state at stage boundaries; that alone removes most of the visible difference. If an order is being chosen for benchmark gain, compare that choice against uniform merging first, because in this regime merging forfeits nothing the best ordering would have bought. Both are cheap, and both are missing from the order-prediction literature, whose validation target is exactly the quantity these two interventions dispose of.

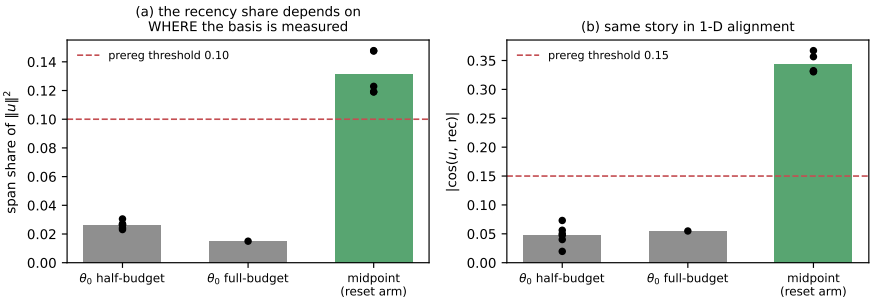

Figure 3: **The basis artifact.** The same u, projected onto single-operator displacement bases measured at θ_0 (half budget and full budget) versus at the phase-boundary midpoints (independent reset runs). Span share (a) and 1-D alignment (b) both cross the preregistered threshold only in the midpoint basis. Dots are seeds.

4 THE PIVOT: THE PARAMETER IMPRINT SURVIVES ALL OF THAT

Section 3 disposed of the benchmark order effect twice over. The natural inference—that there was nothing there—is wrong, and the same reset experiment shows why.

While ΔJ_M collapses, the parameter imprint does not move: $\rho_\theta = \|u_{\text{rst}}\|/\|u\| = 1.007$, with per-seed values $0.982/1.037/1.003$, all inside the preregistered $[0.90, 1.10]$ band that the registration had designated as "optimizer memory is not the main source of the θ-space order effect." It is the same direction too: per-seed $\cos(u_s, u_{\text{rst},s}) = 0.53$–$0.57$, and the reset imprint retains cross-seed consistency ($\bar{\varphi} = 0.472$, baseline 0.494). The function-space channel agrees (order/seed divergence ratio moves $< 30\%$, its preregistered band). Our own registration had predicted $\rho_\theta \in [0.75, 0.95]$; that prediction is falsified on the high side and we report it as such.

This is a dissociation, and it is the reason the rest of the paper works in parameter and function space. A single intervention drives the benchmark quantity below detectability and leaves the parameter quantity numerically unchanged. They cannot be two measurements of one thing. Benchmark deltas are the standard validation target for order-prediction methods; in this setting roughly two-thirds of that delta is Adam bookkeeping, and the part of the composition that reproduces across seeds is invisible to it.

Why care about an imprint benchmarks cannot see? Two reasons the audit later supplies: it is functionally live at finite displacement—moving along it by half its own norm changes J_{avg} measurably while random directions of twice that norm do nothing (§6iii)—and its relative size grows with scale, from $\|u\|/\|d_{AB}\| = 0.97$ at 1.7B to 1.12 at 4B (exploratory, §6v).

5 ACT II: "RECENCY" IS A FUNCTION OF WHERE YOU PUT THE BASIS

The simplest deflationary account of u is recency: each endpoint is dominated by whichever operator ran last, so $u \approx$ (C-displacement) $-$ (M-displacement). Whether the data support this hinges on a free parameter that, to our knowledge, no prior work has controlled: *the point at which the operator displacements defining "recency" are measured* (Figure 1).

θ_0 basis. Using half-budget single-operator displacements $d_M = \theta^{(M)} - \theta_0$, $d_C = \theta^{(C)} - \theta_0$, the span $\{d_M, d_C\}$ explains 2.3–3.1% of $\|u\|^2$ (six seeds), and $|\cos(u, d_C - d_M)| < 0.08$. Full-budget displacements give the same answer (span 1.5%, $|\cos| = 0.055$, seed s_0), so the budget is not the decisive parameter. Taken alone, this reads as a clean rejection of recency, and an earlier version of this project preregistered and "passed" exactly that criterion.

Midpoint basis. But the second operator does not act at θ_0. It acts at the phase-boundary midpoint, and our own function-space measurements show that first-order surrogates transported from θ_0 fail badly in this regime. The optimizer-reset arm of §4 provides exactly the right raw material: $c_{\text{mid}} =$

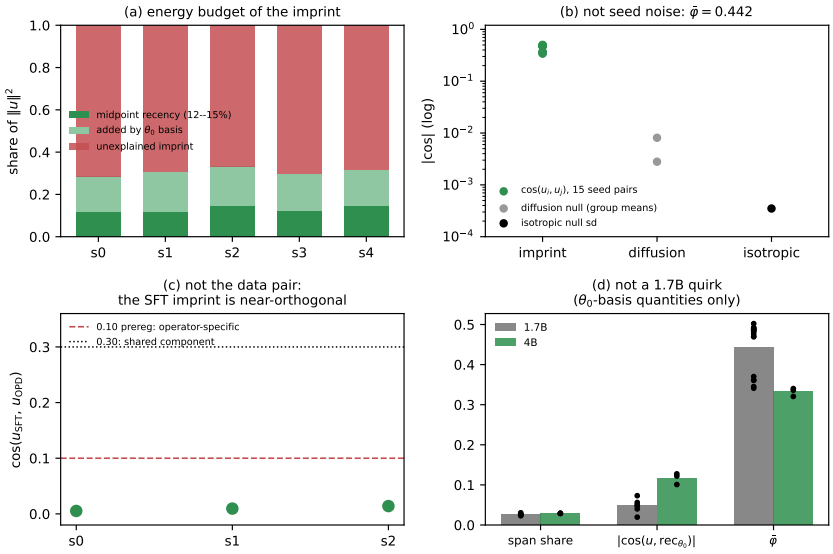

Figure 4: **The energy budget and three nulls.** (a) Energy budget of $\|u\|^2$ per seed after §5. (b) The imprint direction replicates across the 15 seed pairs; matched-seed diffusion nulls and the isotropic null sit two to three orders of magnitude lower (a third diffusion group mean is exactly 0.000 and cannot be shown on a log axis). (c) An SFT pair on the same data leaves a near-orthogonal imprint. (d) θ_0-basis quantities at 1.7B versus 4B; the 4B arm has no reset runs, so midpoint-basis quantities are not measurable there.

$\theta(\text{MC}_{\text{rst}}) - \theta^{(M)}$ is a self-contained realization of "U_C acting at the M-midpoint" (fresh optimizer, real training), and symmetrically m_{mid}. In this basis (five seeds, preregistered before computation): span $\{c_{\text{mid}}, m_{\text{mid}}\}$ explains $f_{\text{mid}} = 11.9$–14.8% of $\|u\|^2$, and $|\cos(u, c_{\text{mid}} - m_{\text{mid}})| = 0.33$–$0.37$—both cleanly across the thresholds that the θ_0 basis sat far below (Figure 3). The four-vector span (midpoint + θ_0 bases jointly) reaches 29–33%.

Function space agrees: the divergence-profile control that survives every calibration in our function-space battery is *compatible* with recency ($D_{\text{order}}/D_{\text{recency}} = 1.28/0.73$), which had appeared to contradict the θ-side rejection. In the midpoint basis the contradiction dissolves: both channels now say recency is part of the story, and disagree only on its share.

We draw a methodological conclusion and a substantive one. Methodological: recency claims about composition are ill-posed until the basis location is stated, and should be reported in at least two bases. Substantive: the last operator's local action accounts for 12–15% of the imprint's energy (29–33% jointly with initialization-basis components)—a real but minority share. One corollary deserves emphasis because it constrains how we read the next section: moving the basis took the unexplained share from 97% to ∼70%, so that share measures how much we looked, not a constant of the model.

6 What the imprint is not, five times over

Roughly 70% of $\|u\|^2$ lies outside every explanatory basis we measured, including the independently retrained midpoint operators. That number is an accounting of our own coverage (§5), so we do not rest anything on it. What the remainder is *not* can be established without it: we try to kill the direction five ways.

(i) Not seed noise. Pairwise $\cos(u_i, u_j)$ across the 15 seed pairs averages $\bar{\varphi} = 0.442$ (excluding the outlier s_5: 0.485; including it, every pair still exceeds 0.34). Matched-seed diffusion nulls (differences of same-order runs across seeds) have mean $|\cos| = 0.0013$ and max group mean 0.0081; the isotropic null has sd 3.5×10^{-4}—two to three orders of magnitude lower, and we report both aggregates. A permutation-alignment control (Git Re-Basin style) finds the identity

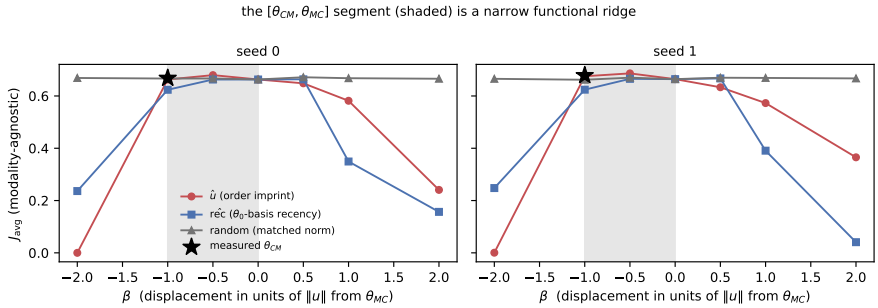

Figure 5: **Functional morphology of the imprint.** J_{avg} along $\theta' = \theta_{MC} + \beta\|u\|\hat{d}$ for $\hat{d} \in \{\hat{u}, \text{rêc}, \text{random}\}$, both seeds, all β (no summary statistic hides the shape). The shaded band marks the $[\theta_{CM}, \theta_{MC}]$ segment for the $\hat{u}$ curve; the star is the independently measured θ_{CM}, which the $\beta{=}-1$ point reproduces (pipeline invariant).

permutation optimal, so φ is not an artifact of neuron relabeling, and a diffusion model of two independent random walks reproduces neither φ nor the norm. The norm statistic survives the harsher aggregate as well: $\|u\|$ exceeds all 30 individual noise-pair norms, though only by a factor of 1.05–1.21. The preregistered $10\times$-max-null line concerns the direction statistic, and there it is met only marginally (9.9–$10.9\times$ across budgets), which we report as such.

(ii) Not the data pair—it is the operator. An SFT pair (same teachers' outputs, same prompts, off-policy cross-entropy) leaves an imprint u_{SFT} nearly orthogonal to the OPD imprint at matched seeds: $\cos(u_{\text{SFT}}, u_{\text{OPD}}) \in [0.005, 0.014]$, 3/3 seeds below the 0.10 line (Figure 4c). The SFT imprint has its own seed consistency, $\varphi_{\text{SFT}} = 0.207$, but it sits at its 0.20 threshold, comes from only three seeds, and 11.6% of the SFT arm's math sequences were trimmed by an OOV truncation unique to that arm, so we treat SFT-side φ as weak evidence. The orthogonality result carries none of those caveats and is the one we lean on: whatever the imprint encodes is a property of the *operator*, not of the data.

(iii) Not an arbitrary high-dimensional direction. We perturb θ_{MC} along matched-norm directions (preregistered; two seeds; Figure 5). Random directions are inert out to $2\|u\|$ (mean $|\Delta J_{\text{avg}}|$ over the $\pm 0.5, \pm 2$ grid is 0.004–0.005). Along $\hat{u}$, the displacement is already measurable at $0.5\|u\|$ ($|\Delta J_{\text{avg}}| = 0.015$–$0.031$); extrapolating to $-2\|u\|$, past θ_{CM}, collapses the model ($J_{\text{avg}} \approx 0$), and $+2\|u\|$ loses 0.30–0.42. The interpolation control completes the picture: the two order endpoints are linearly mode-connected (Garipov et al., 2018; Draxler et al., 2018; Frankle et al., 2020) with no barrier ($B = +0.014/+0.015$ on two seeds, threshold 0.02), and the midpoint sits slightly *above* the endpoints, in the same direction as the merging result of §3. The geometry is thus a single statement: the segment $[\theta_{CM}, \theta_{MC}]$ is a flat plateau on a narrow ridge, equal-norm random directions are inert, and leaving the ridge in either direction along $\hat{u}$ is catastrophic. The recency direction behaves differently: it is flat within $\pm 0.5\|u\|$ ($|\Delta J_{\text{avg}}| = 0.0001$–$0.0027$), and its cliff arrives only at $|\beta| \geq 1$. The imprint's functional response begins at small displacement; recency's does not.

(iv) Not a sharpness or norm artifact. Under diagonal empirical Fisher metrics (Kunstner et al., 2019) estimated at both endpoints and on both domains, $\lambda(u)/\lambda(\text{rand}) = 0.70$–$0.96$ ($4/4$ configurations <1.5; measured at seed s_0 only): the imprint concentrates in directions of *below*-average local curvature. This is consistent with (iii) once both are read as statements about scale: functional specificity lives at finite displacement, not in the local quadratic form. It also answers the sharp/flat reparameterization objection (Dinh et al., 2017) on its own terms, since we report Fisher-metric numbers rather than bare Euclidean norms.

(v) Not a 1.7B quirk. On a Qwen3-4B student (three seeds; restricted to θ-geometry claims by the gate incident of §2), the preregistered criteria replicate under the θ_0-basis qualifier of §5: span share 2.9% (vs. 2–3% at 1.7B), $|\cos(u, \text{rec}_{\theta_0})| = 0.10$–$0.13$ (passes the 0.15 line, fails 0.10—threshold-sensitive, shown with the scan), and $\bar{\varphi} = 0.332$ with all three pairs above 0.32. The 4B arm has

no reset runs, so the midpoint-basis decomposition is not measurable there; exploratorily, the two composition paths are less parallel at 4B ($\cos(d_{AB}, d_{BA})$ drops from 0.53 to 0.37) and the relative imprint grows ($\|u\|/\|d_{AB}\|$: $0.97 \to 1.12$).

Exploratory: not a token-level object either. 27.8–38.8% of $\|u\|^2$ across the six seeds (38.6% at s_0; the low end is s_5) sits in the shared embedding, inviting a token-level reading. It fails: the top-500 rows by norm carry only 1.1–1.5% of the embedding energy (diffuse, not compact); their character-class mix matches a random draw except that code symbols are *depleted*; top rows are low-frequency subword fragments. Yet row support is far from random across seeds (Jaccard 0.08–0.32 vs. random expectation 0.0016; s_5 lowest). The imprint is inconsistent with "specific knowledge written to specific tokens" and consistent with a systematic, diffuse path signature of the training dynamics. This paragraph is exploratory: no preregistered criterion governs it.

7 The channel we wanted and could not measure

The occupancy share is the quantity that §1 argues makes on-policy composition distinct—order effects flowing through *which states the policy visits*—and it is the one number this paper does not have. We report the attempt in full, because the variance budget is what the next attempt needs.

Our finite-difference bracket estimator produced a seductive single-point answer (normalized overlap 0.887, SNR 2237) that an ε-scan exposed as bf16 rounding noise: the estimate scaled as $1/\varepsilon$. The replacement score-function estimator then failed its preregistered self-distillation calibration gate, but a preregistered disjoint-half adjudication (Appendix B.2) shows that the estimator is *unbiased* and the occupancy path in the code intact, so the obstacle is purely statistical. Per-sequence noise is $\sim 40\times$ the signal; a token-level Rao–Blackwellized variant bought only a 1.2–$1.4\times$ norm reduction (the preregistration required $\geq 3\times$); and passing the calibration gate would take $\sim 6 \times 10^4$ sequences per direction, far beyond our budget but not beyond a well-resourced follow-up. We report no occupancy number, and we note that the residual of §6 is where such a term would have to live.

The lesson generalizes past this estimator: a single-point estimate with an impressive SNR against the *wrong* noise floor is how plausible false results are manufactured. Our noise floor at the time measured the repeatability of recomputing the same rollouts, which is not the error source that finite differencing suffers from.

Scope not covered. Budget-transfer comparisons of function-space divergences across T are absolute-value comparisons we did not calibrate and do not report. Step-wise trajectory norms were not logged in the main matrix. A second task pair does not exist: its general-chat operator destroyed spontaneous termination and the arm was terminated at its gate (Table 1), so we claim nothing beyond math$\times$code. All three are open instrumentation, not evidence of absence.

8 Related work

Order prediction via brackets. That training sequence matters is old (curriculum learning, Bengio et al., 2009; catastrophic forgetting, Kirkpatrick et al., 2017); predicting it is recent. Rukhovich et al. (2024) introduced Lie-bracket analysis of training-order choice for bilingual pretraining; Sweeney (2026a) scaled bracket-score tournaments to instruction-SFT/DPO ordering; Piontkovskaia & Nikolenko (2026) show two-step task-gradient order sensitivity is first-order predictable with the Hessian-only bracket $H_B g_A - H_A g_B$. All three predict *which* order is better; none measures what the order difference is made of, and their fixed-dataset settings structurally lack the occupancy channel that makes on-policy composition distinct (their leading term confirms this: with θ-independent sampling the occupancy bracket vanishes identically). Our results also complicate their validation target: the benchmark delta they predict is, in our setting, roughly two-thirds optimizer memory (§3) and fully replicated by uniform merging, while the component that reproduces across seeds is not visible in it at all. We confirm the single-task mechanism of Sweeney (2026b) at the cross-operator level.

Table 2: Composition-claim checklist. Each row is a control this paper found load-bearing, and what happened in its absence.

If you claim	Control it needs	What we measured without it
order A beats order B on a benchmark	optimizer state reset at the stage boundary	$\sim$2/3 of ΔJ_M is Adam moments (§3)
order choice is worth making	uniform weight averaging as a baseline	merging matches the better order (§3)
the difference is / is not recency	≥ 2 basis locations (θ_0 and phase boundary)	span share $3\% \to 15\%$ on the same u (§5)
the difference is a real direction	matched-seed nulls + permutation alignment	nulls 2–3 orders below $\bar{\varphi}$ (§6i)
the effect is operator interaction	a same-data, different-operator arm	SFT imprint near-orthogonal (§6ii)
any estimator with a free step size	scan ≥ 2 values of it	SNR 2237 was bf16 rounding (§7)
any operator in the composition	preregistered own-domain gain gate	4 of 5 configurations failed (§2)

Parameter vs. function distance. That parameter distance poorly proxies function change is long known (Benjamin et al., 2019), as is norm growth with directional convergence (Soudry et al., 2018; Hoffer et al., 2017) and the sharpness reparameterization critique (Dinh et al., 2017); gradient descent confining itself to low-dimensional subspaces (Gur-Ari et al., 2018) motivates our span tests. We differ in measuring *paired order differences* with matched seed nulls rather than single-trajectory drift, and in reporting Fisher-metric invariants (Wang et al., 2026) alongside Euclidean ones.

Merging and connectivity. Weight averaging (Wortsman et al., 2022), task arithmetic (Ilharco et al., 2023) and its tangent-space analysis (Ortiz-Jiménez et al., 2023), interference-aware merging (Matena & Raffel, 2021; Yadav et al., 2023; Yu et al., 2024), and permutation-aligned mode connectivity (Ainsworth et al., 2023; Entezari et al., 2021) supply our order-free baselines and the relabeling null; Zhou & Xie (2026) argue capabilities live in activation subspaces rather than weight geometry, consonant with our function-space channel. Model-diffing work (Minder et al., 2025b;a; Kempf et al., 2026) compares two models in activation space; we audit a *difference of differences* in weight space with seed replication. Kuditipudi et al. (2025) detect data-order traces behaviorally, a precedent for order leaving recoverable marks. Seed and data-order variance as a baseline hazard goes back to Dodge et al. (2020).

9 Limitations and conclusion

Limitations. (1) One task pair: every quantitative claim is about math$\times$code OPD on Qwen students, and the gate failures of §2 show extension is not automatic. (2) The benchmark-level order effect we dissociate is itself modest, and the optimizer-reset reduction is significant as a level (post-reset residual indistinguishable from zero) but not as a paired difference at five seeds (§3). (3) The SFT comparison arm carries scoring caveats (§6ii). (4) 4B replication is θ-geometry only, under the θ_0-basis qualifier. (5) The unexplained share is a share of *measured* bases; §5 is a worked demonstration that it shrinks when more is measured, which is what an audit residual means. (6) Absolute bf16-scale quantities are reported only as ratios against matched nulls. (7) Fisher and permutation controls are single-seed.

Conclusion. "The order of post-training matters" bundles two claims that one experiment pulls apart. The benchmark order effect is mostly optimizer state crossing a stage boundary, and what survives that is matched for free by averaging the two single-operator models, so order is not a lever for benchmark gain. The parameter order imprint is untouched by the same intervention: a seed-replicable, operator-specific, functionally narrow, Fisher-flat direction whose recency share depends on an unstated basis choice and whose remainder survives every null we could construct and one change of scale. The literature validates order predictions against the first quantity; the second is the one that reproduces. Composition claims, predictive or mechanistic, should be audited at least as hard as Table 2 before they are believed.

AI USE STATEMENT

In this work, we used generative AI tools for two purposes, both in the recommended disclosure category: (i) editing the paper for readability (grammar, wording, and concision of author-written text), and (ii) reviewing software code, i.e., reading the training, evaluation, and scoring code and flagging potential defects. We have not used generative AI tools to generate synthetic data, develop theoretical frameworks, formulate mathematical claims or proofs, propose or refine hypotheses, design experiments, implement methods, clean datasets, analyze data, or interpret results; translation is not applicable to this work. We have reviewed all AI-assisted work: every suggested text edit was checked by the authors against the intended technical meaning, and every defect flagged during code review was independently reproduced by the authors before any fix was made, with affected numbers recomputed from saved raw completions. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All preregistrations (criteria frozen before computation), change ledgers, per-seed numbers, evaluation completions, figure scripts, and analysis scripts will be released. Every figure value traces to a JSON artifact named in the appendix claim–evidence table (Table 3).

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2306.13649.

Samuel K. Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2209.04836.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, pp. 41–48, 2009. doi:10.1145/1553374.1553380.

Ari S. Benjamin, David Rolnick, and Konrad Kording. Measuring and regularizing networks in function space. In *International Conference on Learning Representations (ICLR)*, 2019. arXiv:1805.08289.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, et al. Accounting for variance in machine learning benchmarks. *arXiv preprint arXiv:2103.03098*, 2021.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

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

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*, 2023.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning (ICML)*, 2017. arXiv:1703.04933.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*, 2020.

Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred A. Hamprecht. Essentially no barriers in neural network energy landscape. In *International Conference on Machine Learning (ICML)*, pp. 1308–1317, 2018. PMLR 80; arXiv:1803.00885.

Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. The role of permutation invariance in linear mode connectivity of neural networks. *arXiv preprint arXiv:2110.06296*, 2021.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M. Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *International Conference on Machine Learning (ICML)*, 2020. arXiv:1912.05671.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry Vetrov, and Andrew Gordon Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. arXiv:1802.10026.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: On-policy distillation of large language models. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2306.08543.

Guy Gur-Ari, Daniel A. Roberts, and Ethan Dyer. Gradient descent happens in a tiny subspace. *arXiv preprint arXiv:1812.04754*, 2018.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. arXiv:2103.03874.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. NIPS 2014 Deep Learning Workshop.

Elad Hoffer, Itay Hubara, and Daniel Soudry. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. arXiv:1705.08741.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, et al. Qwen2.5-coder technical report. *arXiv preprint arXiv:2409.12186*, 2024.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2212.04089.

Elias Kempf, Simon Schrodi, Bartosz Cywiński, Thomas Brox, Neel Nanda, and Arthur Conmy. Simple llm baselines are competitive for model diffing. *arXiv preprint arXiv:2602.10371*, 2026.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 2017. arXiv:1612.00796; doi:10.1073/pnas.1611835114.

Rohith Kuditipudi, Jing Huang, Sally Zhu, Diyi Yang, Christopher Potts, and Percy Liang. Blackbox model provenance via palimpsestic membership inference. *arXiv preprint arXiv:2510.19796*, 2025.

Frederik Kunstner, Lukas Balles, and Philipp Hennig. Limitations of the empirical fisher approximation for natural gradient descent. *arXiv preprint arXiv:1905.12558*, 2019.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *ACM SIGOPS Symposium on Operating Systems Principles (SOSP)*, 2023. arXiv:2309.06180.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, et al. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

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

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. *arXiv preprint arXiv:2305.01210*, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. arXiv:1711.05101.

Michael Matena and Colin Raffel. Merging models with fisher-weighted averaging. *arXiv preprint arXiv:2111.09832*, 2021.

Julian Minder, Clément Dumas, Caden Juang, Bilal Chugtai, and Neel Nanda. Overcoming sparsity artifacts in crosscoders to interpret chat-tuning. *arXiv preprint arXiv:2504.02922*, 2025a. NeurIPS 2025.

Julian Minder, Clément Dumas, Stewart Slocum, Helena Casademunt, Cameron Holmes, Robert West, and Neel Nanda. Narrow finetuning leaves clearly readable traces in activation differences. *arXiv preprint arXiv:2510.13900*, 2025b. Accepted to ICLR 2026.

Brian A. Nosek, Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606, 2018. doi:10.1073/pnas.1708274114.

Guillermo Ortiz-Jiménez, Alessandro Favero, and Pascal Frossard. Task arithmetic in the tangent space: Improved editing of pre-trained models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2305.12827.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

Irina Piontkovskaia and Sergey Nikolenko. First-order predictable but pairwise fragile: Local task adaptation in trained transformers. *arXiv preprint arXiv:2607.16821*, 2026.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023.

Alexey Rukhovich, Alexander Podolskiy, and Irina Piontkovskaya. Commute your domains: Trajectory optimality criterion for multi-domain learning. In *NeurIPS Workshop on Mathematics of Modern Machine Learning (M3L)*, 2024. arXiv:2501.15556.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70): 1–57, 2018. arXiv:1710.10345.

Gilbert Strang. On the construction and comparison of difference schemes. *SIAM Journal on Numerical Analysis*, 5(3):506–517, 1968. doi:10.1137/0705041.

John Sweeney. The geometry of sequential learning: Lie-bracket prediction of transfer order. In *International Conference on Machine Learning (ICML)*, 2026a. arXiv:2606.24993.

John Sweeney. Optimizer memory makes shuffle order a first-order source of fine-tuning noise. *arXiv preprint arXiv:2606.29554*, 2026b.

Jiayu Wang, Zuojun Ye, and Wenpeng Yin. Functionality-oriented llm merging on the fisher–rao manifold. *arXiv preprint arXiv:2603.04972*, 2026.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning (ICML)*, 2022. arXiv:2203.05482.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2306.01708.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, et al. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2311.03099.

Rui Zhou and Tianci Xie. Fora: From weight-space to function-space protection in capability-preserving fine-tuning. *arXiv preprint arXiv:2606.31092*, 2026.

A Claim–evidence table

Every headline claim, its preregistration (criteria frozen before the deciding computation), and the artifact its numbers are read from. Internal document numbers (D–) refer to the released project ledger; artifact names are the released JSON products. Claims marked E are exploratory (no preregistered criterion).

B Additional experimental detail

B.1 The optimizer-reset comparison, in full

Section 3 quotes a matched-seed shrinkage. The underlying per-seed values are below. The preregistered criterion (docs A1 §E3, frozen before any reset checkpoint existed) is on the *level* $|\Delta J_M^{\mathrm{rst}}| \leq 0.006$, not on a ratio; the reset mean -0.0038 (sd 0.0061, se 0.0027) clears it.

B.2 Occupancy estimator: the disjoint-half unbiasedness adjudication

The score-function occupancy estimator failed its self-distillation calibration gate, which leaves two possibilities: a biased estimator (a bug) or an unbiased one drowned in variance. A preregistered follow-up separated them by disjoint-half cross products. Three mutually independent unbiased estimates of the signal power—$\langle Hv_1, Hv_2\rangle$, $\langle Ov_1, Ov_2\rangle$, and $-\langle Hv, Ov\rangle$ across halves—agree with the predicted $\|Fv\|^2 \approx 350$ under the frozen same-sign, same-order-of-magnitude criterion: the Hessian-side estimate lands at 332 and 426 for the two probe directions, and the cross-term-to-Hessian ratios are 0.49 and 1.09, inside the frozen $[1/3, 3]$ band. The noisier O-side estimates scatter up to 2.7×10^3, which is itself the variance diagnosis. The estimator is therefore unbiased and the obstacle statistical, as §7 reports.

B.3 Generation-config parity (training vs. evaluation)

Both sides drive the same vLLM engine (bf16; Kwon et al., 2023) with an explicit `stop_token_ids` list and the same chat-template builder; the shared token budget constant is 1024. The one intentional difference is temperature: training rollouts sample at $T{=}1.0$ (on-policy distillation needs draws from the current policy), evaluation is greedy.

Table 3: Claim–evidence–preregistration map.

Claim (section)	Prereg	Status	Artifact
$\Delta J_M = -0.0146 \pm 0.0053$, 5/6 signs; ΔJ_{avg} 2.4 se (§3)	D08	adjudicated	`order_effect_modality_6seed.json`
Optimizer reset: $-0.0107 \to -0.0038$ matched-seed, residual passes $\lvert\Delta J_M\rvert \leq 0.006$ (§3)	D14	adjudicated	E4 eval set + `theta_geom_t400_rst.json`
$\rho_\theta = 1.007$; $\cos(u, u_{\text{rst}}) \in [0.53, 0.57]$ (§4)	D14/D24	adjudicated	`theta_geom_t400_rst.json`, `midpoint_recency.json`
θ_0-basis span 2.3–3.1%, $\lvert\cos\rvert < 0.08$ (§5)	D12	adjudicated	`theta_geom_t400_s012345.json`
Full-budget θ_0 basis: span 1.5% (§5)	D24	descriptive	`theta_geom_fullref_s0.json`
Midpoint basis: $f_{\text{mid}} = 11.9$–14.8%; 4-basis 29–33% (§5)	D24	adjudicated	`midpoint_recency.json`
Function-space recency control $D_{\text{ord}}/D_{\text{rec}} = 1.28/0.73$ (§5)	D13	adjudicated	E3 product set
$\lambda = 0.5$ merge $\approx$ better order/joint; $\lambda = 1.0$ worse (§3)	declared	replicated	TA eval set
No barrier on $[\theta_{CM}, \theta_{MC}]$: $B = 0.014/0.015$ (§6iii)	D21	adjudicated	`pert_u_s{0,1}_7b.json`
$\bar{\varphi} = 0.442$; nulls 2–3 orders lower (§6i)	D12/D19	adjudicated	`theta_geom_t400_s012345.json`
$\lVert u\rVert >$ all 30 noise pairs (1.05–1.21×); 10× line marginal on $\bar{\varphi}$ (§6i)	D12/D07	borderline	same
Permutation null: identity optimal (§6i)	D19	adjudicated	`n2_permutation_s0.json`
SFT/OPD orthogonality $\cos \leq 0.014$ (§6ii)	D18	adjudicated	`theta_geom_t400_sft.json`, `s2d_cross_arm.json`
Ridge morphology: random inert; extrapolation collapse (§6iii)	D21	adjudicated	`pert_{u,rand,rec}_s{0,1}_7b.json`
Fisher-flat: $\lambda(u)/\lambda(\text{rand}) = 0.70$–$0.96$, 4/4, seed s_0 (§6iv)	D21	adjudicated	`fisher_*.json`
4B replication with threshold scan (§6v)	D17	adjudicated	`theta_geom_4b.json`
Embedding dissection: diffuse, domain-agnostic, shared support (§6)	E	exploratory	`u_embedding_tokens.json`
Gate history (Table 1)	D05/D17/D23	adjudicated	gate eval set
Occupancy share: estimator validated unbiased, share unmeasured (§7)	D11	retired	`commutator_*.json`, `occupancy_stage0_s0.json`

B.4 Truncation rates

Fraction of completions that hit the 1024-token cap (`finish_reason=length`), $T=400$, 7B-teacher arm, range over seeds:

B.5 Task-arithmetic per-seed values

Modality-agnostic J_{avg} (and J_M/J_C for $\lambda=0.5$), per seed:

B.6 Ridge-morphology grid (data behind Figure 5)

$J_{\text{avg}}(\theta^s_{MC} + \beta \lVert u\rVert \hat{d})$ on the full β grid; $\beta=0$ is θ_{MC} itself and $\hat{d} = \hat{u}$ at $\beta = -1$ lands on θ_{CM} up to evaluation noise (measured gap $-0.005/-0.003$).

seed	s_0	s_1	s_2	s_3	s_4	s_5	mean $\pm$ se
baseline ΔJ_M	$-.0191$	$-.0180$	$-.0128$	$+.0048$	$-.0084$	$-.0342$	$-.0146 \pm .0053$
(matched 5)	$-.0191$	$-.0180$	$-.0128$	$+.0048$	$-.0084$	—	$-.0107 \pm .0043$
reset ΔJ_M	$-.0060$	$-.0025$	$-.0123$	$-.0030$	$+.0046$	—	$-.0038 \pm .0027$
paired (rst$-$base)	$+.0131$	$+.0155$	$+.0005$	$-.0078$	$+.0130$	—	$+.0069 \pm .0045$

Table 4: Per-seed accuracy deltas for the optimizer-memory experiment. Three readings of the same table, all reported in §3: the level after reset is indistinguishable from zero (1.4 se) and passes the frozen line; the matched-seed shrinkage is $0.36\times$ ($0.42\times$ on the three seeds the criterion was registered on, $-.0166 \to -.0069$); and the paired reduction is 1.5 se and non-uniform, vanishing at s_2 and reversing at s_3. The reset arm was not run at s_5. Earlier project notes quoted $0.23\times$, which divides the five-seed reset mean by the three-seed baseline mean; we quote the matched comparison instead.

	training rollouts	evaluation
backend / dtype	vLLM, bf16	vLLM, bf16
temperature	1.0 (sampled)	0.0 (greedy; intentional)
max new tokens	1024	1024 (same constant)
stop tokens	{<|im_end|>, eos}	{<|im_end|>, eos, pad}
prompt template	`build_chat_prompt`	same function
artifacts	—	raw completions always saved

Table 5: Generation-config parity. An earlier version of the evaluation harness relied on the model's own `generation_config` eos (for the Qwen3 base that is <|endoftext|>); distilled checkpoints learn to emit <|im_end|> instead, so they were forced to the token cap while the untrained baseline stopped normally—an asymmetry that penalized exactly the models under study. It was caught and fixed before any number in this paper was produced; we list it as one more instance of the parity failures that motivate §2.

B.7 4B REPLICATION INVENTORY

The 4B arm (Qwen3-4B-Base student, same 7B teachers) evaluated the full θ-geometry matrix—$\{M_{\text{half}}, C_{\text{half}}, \theta_{MC}, \theta_{CM}, \text{joint}\} \times 3$ seeds—plus the base model and single-operator gate probes (M-only, C-only, seed s0). Per the scope reduction registered after the gate failure (§2), the 4B arm supports *θ-geometry claims only* (span, φ, $\cos(d_{AB}, d_{BA})$, R); no J-level order-effect claim is made at 4B, and no conditions beyond the fifteen matrix runs were trained.

condition	GSM8K	HumanEval$^+$
π_0	0.5%	0.6%
M_{half}	100%	12–23% (s5: 100%)
C_{half}	19–94%	100%
θ_{MC} (M→C)	1–13% (s5: 100%)	100%
θ_{CM} (C→M)	100%	100%
joint / alt. / Strang (Strang, 1968)	100%	100%
TA λ=0.5	∼1% (s5: 36%)	100%
TA λ=1.0	89–100%	100%

Table 6: Cap rates. Three facts keep this from contaminating the headline numbers. (i) It is real model behavior, not an evaluation-config mismatch: a same-vintage pair of runs (θ_{MC} vs. θ_{CM}, seed s1, identical harness) still differs $4\times$ in mean output length. (ii) The answer arrives *before* the cap and degenerate repetition follows (missing-answer rates stay at fractions of a percent), which is what the modality-agnostic scoring rule absorbs: it extracts the first well-formed answer. (iii) The absorption was measured, not assumed: raw and modality-agnostic scores agree on every condition except C_{half}'s J_M (0.579→0.632 at T=400, 0.558→0.602 at T=500). Two behavioral observations are worth recording as exploratory: the seed-s5 outlier of §2 is visible here as a stopping-behavior signature (100% cap where the other θ_{MC} seeds stop normally), and merging at λ=0.5 restores normal stopping on math while λ=1.0 does not.

		TA λ=0.5		TA λ=1.0	θ_{MC}	θ_{CM}
seed	J_M	J_C	J_{avg}	J_{avg}	J_{avg}	J_{avg}
s0	.7398	.6265	.6831	.6679	.6631	.6682
s1	.7276	.6116	.6696	.6612	.6644	.6785
s2	.7327	.6265	.6796	.6641	.6656	.6703
s3	.7350	.6300	.6825	.6652	.6676	.6724
s4	.7256	.6291	.6774	.6698	.6743	.6717
s5	.7137	.6287	.6712	.6576	.6657	.6706
T=500 s0–s2	.7319–.7406	.6240–.6371	.6823–.6845	.6579–.6683		

Table 7: Task arithmetic per seed. "Better order" in §3 is θ_{CM}, which wins on the six-seed mean; the paired difference λ=0.5 minus θ_{CM} is $+0.0053 \pm 0.0034$ (5/6 positive, s1 the exception). Against the per-seed maximum of the two orders—a harsher comparator—it is $+0.0049 \pm 0.0035$, the same verdict. The T=500 row is shown for completeness.

seed	$\hat{d}$	β=−2	−1	−0.5	+0.5	+1	+2
s0	$\hat{u}$	.000	.663	.680	.648	.582	.241
s0	rec	.236	.624	.663	.664	.349	.157
s0	rand	.669	.667	.667	.672	.668	.666
s1	$\hat{u}$	.001	.676	.686	.633	.573	.365
s1	rec	.248	.624	.666	.667	.391	.041
s1	rand	.666	.662	.670	.670	.669	.667

Table 8: β=0 base points: $J_{\text{avg}}(\theta_{MC}) = .663/.664$, $J_{\text{avg}}(\theta_{CM}) = .668/.678$ (s0/s1).