GROUP-RESOLVED WEIGHT DECAY CONTROLS GROKKING ARRIVAL ACROSS CHECKPOINT STATES

Anonymous authors
Paper under double-blind review

ABSTRACT

Does grokking depend only on how much weight decay is applied, or also on where it is applied? We study two-layer Transformers using paired continuations from complete checkpoint states. We allocate decay between internal computation parameters (attention, MLP, LayerNorm; GC) and input/output parameters (embedding, head; EH), matching initial decay-only squared-norm contraction. On modular addition, shifting this matched total from EH toward GC advances sustained high test accuracy by $2{,}100$ steps on average, with the same direction in all 12 parents. Both groups accelerate arrival, but GC has the larger response in the tested dose units. At early checkpoints, high-GC continuations arrive in $63/64$ cases within $8{,}000$ steps, versus $0/64$ under low GC. Extending all 128 low- and intermediate-GC arms to $32{,}000$ steps yields arrival in every arm: the short-budget gap reflects delay for these runs, not permanent failure. The fixed allocation swap also advances arrival on S_5 permutation composition. Separate parameter-source interventions and a one-step response assay characterize complementary effects of inherited state, without establishing a mechanism for arrival. These results identify decay allocation as a control of grokking timing and show why finite-budget completion must be distinguished from eventual generalization.

1 INTRODUCTION

Grokking is delayed generalization after a model has already fit its training set (Power et al., 2022). Explanations emphasize circuit formation and cleanup (Nanda et al., 2023; Zhong et al., 2023), circuit-efficiency competition (Varma et al., 2023), representation and norm dynamics (Liu et al., 2022; 2023), and optimizer dynamics (Thilak et al., 2022). We ask two linked questions: which parameter groups respond more strongly to a specified decay intervention, and when does faster arrival appear as a difference in whether generalization occurs at all? Answering the second question requires varying both the checkpoint state and the observation horizon.

Separating the effects of parameter groups. Uniform weight decay changes all parameter groups together. To separate their effects, we divide the model into the core group (GC: attention, MLP, and LayerNorm parameters, i.e. the internal computation parameters) and the embedding/head group (EH: input embeddings and the output head, i.e. the input/output parameters). GC groups hidden-state transformations with their normalization, including final LayerNorm; EH contains the input and output interfaces. These distinct architectural roles motivate separating decay, but do not by themselves predict which allocation will advance generalization more. We independently vary decay on the two groups, restoring the same weights, Adam moments, batch stream, and random-number state before each continuation. Group-selective decay has also been studied as an optimization strategy (He et al., 2025); here it provides a controlled measurement of generalization timing.

Comparing interventions from the same checkpoint. We choose decay strengths using $q_g = \lambda_g \|\theta_g(t_{\mathrm{fork}})\|_2^2$, which measures initial squared-norm contraction due to decay. At the primary checkpoint, equal decay coefficients would give EH about $2.6\times$ the contraction mass of GC. Matching q_g gives a common initial contraction scale, not equal cumulative decay or functional perturbation. We compare allocations, checkpoint stages, and observation horizons, then test a fixed swap on S_5 (Figure 1). Our aim is to identify allocation effects under controlled checkpoint interventions.

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.

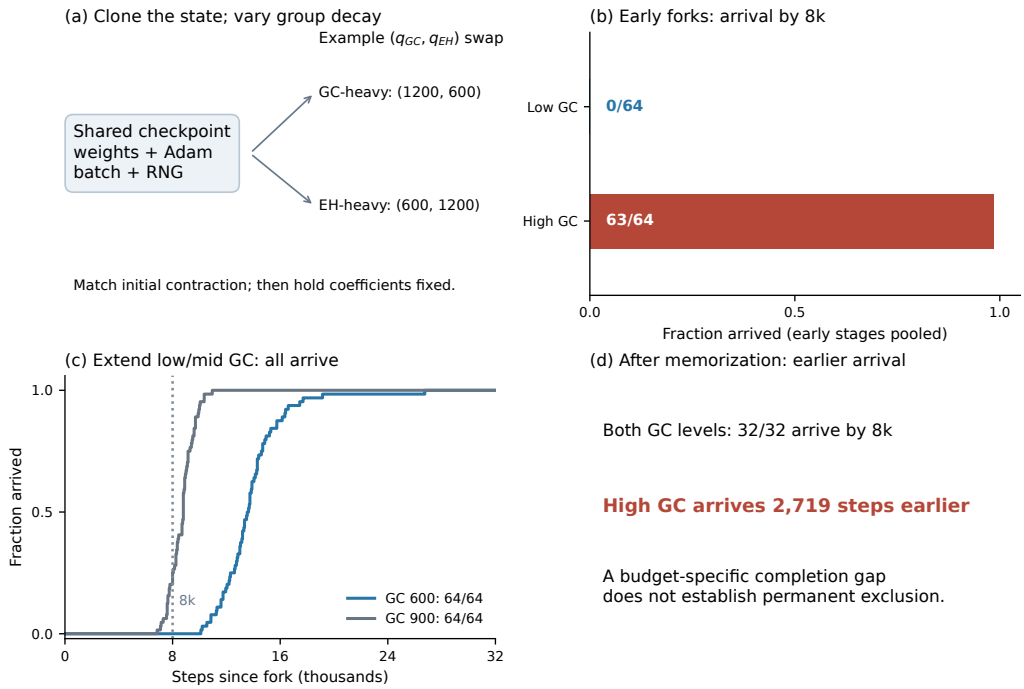


Figure 1: From a finite-budget gap to observed delay. (a) Full-state cloning isolates group-decay interventions; the example swap is selected through initial pressure. (b) Early forks show a large 8k completion gap. (c) All 128 extended low-/mid-GC arms arrive by 32k; high-GC arms were not extended. (d) After memorization, both GC levels complete by 8k, with high GC arriving earlier. Counts pool arms; inference uses the 16 paired parents.

Two principal findings.

1. **Allocation changes arrival.** At the same total initial pressure, shifting decay toward GC advances modular-addition arrival by 2,100 steps (12/12 parents). A factorial separates the positive responses of both groups.
2. **A completion gap can be a timing gap.** Paired stage comparisons and extended observation show how a near-complete early 8k separation becomes eventual arrival in all 128 extended low-/mid-GC arms.

Supporting tests. A source-fixed allocation swap on S_5 tests the first finding on a second task. Separate parameter-source and one-step response experiments examine inherited state at behavioral and local-response timescales; they do not identify a geometry-to-arrival mechanism.

2 Related Work

Representations and learning dynamics. Mechanistic studies identify circuits and representations underlying grokking (Nanda et al., 2023; Zhong et al., 2023; Varma et al., 2023; Liu et al., 2022). Gromov (2023) gives analytic modular-arithmetic solutions and demonstrates grokking without regularization in two-layer networks; Stander et al. (2024) identify subgroup-based circuits for S_5 and S_6. Omnigrok connects generalization to loss landscapes (Liu et al., 2023), while Kumar et al. (2024) describe delayed feature learning through a transition from lazy to rich dynamics. These accounts motivate examining how interventions change the time required to reach a generalizing solution.

Norms, optimization, and timing. Slingshot links delayed generalization to adaptive-optimizer dynamics (Thilak et al., 2022); Prieto et al. (2025) study logit scaling and numerical stability, and Boursier et al. (2025) derive slow norm-minimizing dynamics under a bounded-flow assump-

tion. Recent norm-delay and decay-clock studies provide further timing descriptions (Truong et al., 2026a;b; Kim, 2026; Verma, 2026). Checkpoint chimeras and norm-to-logit interventions (Truong, 2026c;d) motivate our paired design. We separate future group-decay interventions from the choice of inherited parameter and optimizer sources, measuring both arrival and immediate response.

Parameter-group optimization and regularization. LARS and LAMB adapt update scales across layers (You et al., 2017; 2020); AlphaDecay allocates weight decay across modules (He et al., 2025). Embedding-gradient studies also motivate examining groups separately (AlquBoj et al., 2025). These approaches establish a broader precedent for group-dependent optimization. With AdamW's decoupled decay (Loshchilov and Hutter, 2019), our intervention holds the learning rate fixed and varies group decay from matched checkpoint states. Rather than asking whether a group-dependent optimizer improves training, we ask whether allocation changes grokking time at a matched initial contraction scale. Complete-state pairing isolates the continuation-policy contrast; extended observation then tests whether its finite-budget completion gap persists. These are distinct questions from identifying a norm-based clock or a circuit responsible for generalization.

3 Paired Interventions and Arrival Measurement

A *parent* is an independently trained source run. A *fork* is the checkpoint at which intervention begins; an *arm* is one continuation from that checkpoint. A cohort groups the parents used in an experiment.

3.1 Task, model, and checkpoint controls

The primary task is addition over $\mathbb{Z}_{97}$ with a deterministic 30% training, 20% calibration, and 50% final-test split. The model is a two-layer causal Transformer (width 128, four heads, feed-forward width 512, zero dropout). AdamW uses learning rate 10^{-3}, betas $(0.9, 0.999)$, and batch size 512. The GC/EH grouping follows the introduction; Appendix B gives eligibility and implementation details.

The primary factorial uses 12 eligible step-800 parents trained under uniform decay. Each arm restores model weights, optimizer moments, batch-generator state, and RNG state before changing treatment. The separate three-stage cohort uses 16 fresh parents with event-relative forks; the transplant instead deliberately assembles distinct source components. Parents are the inferential units. Paired sign-flip tests and parent resampling retain within-parent dependence.

3.2 Matching the initial decay intervention

For parent p at fork t_f, define

$$S_{p,g} = \|\theta_{p,g}(t_f)\|_2^2, \qquad \lambda_{p,g}(q_g) = \frac{q_g}{S_{p,g}}, \qquad (q_{\mathrm{GC}}, q_{\mathrm{EH}}) \in \{600, 900, 1200\}^2. \tag{1}$$

Coefficients are computed once and held fixed throughout continuation. For the decay-only update, the exact squared-norm contraction is $2\eta q_g - \eta^2 q_g^2 / S_{p,g}$; its first-order value is $2\eta q_g$. At the primary fork, $S_{\mathrm{GC}} \approx 989$ and $S_{\mathrm{EH}} \approx 2595$: equal coefficients deliver unequal contraction masses. The grid yields 108 continuations matched on this initial quantity.

We call q the initial *pressure*: it measures the decay-only Euclidean contraction, not the total optimizer update. This interpretation follows from AdamW's decoupling of decay and adaptive gradient scaling (Loshchilov and Hutter, 2019). Pairwise contrasts estimate the total effect of these fixed continuation policies selected at the fork.

3.3 Measuring arrival within a fixed budget

Arrival requires train accuracy ≥ 0.99 and test accuracy ≥ 0.90 at five consecutive evaluations. Onset t_o is the episode's first checkpoint; confirmation is $t_c = t_o + 200$ (50-step spacing). We retain implemented timestamps rather than harmonizing them retrospectively: changing the clock can alter censored contrasts.

Table 1: Evidence map. Each row summarizes a separate experiment.

Question	Supported finding	Experiment
Group allocation	Equal-total swap advances arrival by 2,100 steps	12-parent factorial; both group responses positive
State and budget	Early completion gap resolves as delay	16-parent stage comparison; 128 extended low-/mid-GC arms
Task transport	S_5 advance of 1,063 steps at 12k	12-parent source-fixed allocation swap
Parameter source	Arrival advance of 810 steps	32-parent parameter/optimizer-source factorial
Local response	Jacobian term dominates signed source contrast	Independent 24-parent one-step assay; not arrival mediation

The primary factorial and transplant score onset; stage/refork cohorts score confirmation. The S_5 analysis scores onset only if confirmation occurs by the cutoff. Appendix C gives the cohort-specific rules and boundary cases.

For fork t_f and budget L, let $I_{p,L}$ indicate arrival under the cohort's observation rule and r_p its elapsed timestamp ($t_o - t_f$ or $t_c - t_f$). Define

$$Y_{p,L} = \begin{cases} r_p, & I_{p,L} = 1, \\ L, & I_{p,L} = 0, \end{cases} \qquad \Gamma_{H,p,L} = 1 - Y_{p,L}/L. \tag{2}$$

Thus Y is elapsed time capped at the budget, while Γ_H rescales it so that earlier arrival scores higher. Unobserved arrival contributes the bounded value L. The primary $L = 11{,}200$ corresponds to calendar cutoff $C_H = 12{,}000$ at $t_f = 800$; the stage cohort uses $L = 8{,}000$. Calendar scores satisfy $Y_H^{\mathrm{cal}} = t_f + Y_{p,L}$, so paired calendar and elapsed contrasts agree within each parent. Onset and confirmation scores can differ in contrasts involving censoring.

For example, the equal-total allocation estimand is the parent-average of

$$D_{p,L} = Y_{p,L}(600, 1200) - Y_{p,L}(1200, 600), \tag{3}$$

positive when GC-heavy arrives earlier. This compares two fixed-coefficient continuation policies selected through fork-time pressure. Group main effects average high-minus-low score differences over the other group's levels. With $x_g = (q_g - 900)/300$, we summarize the grid by $\Gamma_H = a_p + w_{H,\mathrm{GC}} x_{\mathrm{GC}} + w_{H,\mathrm{EH}} x_{\mathrm{EH}} + \varepsilon$. These coefficients summarize additive responses over the tested grid.

4 Group-Resolved Arrival Control

Allocation matters even at the same total initial pressure. The GC-heavy policy $(q_{\mathrm{GC}}, q_{\mathrm{EH}}) = (1200, 600)$ arrives 2,100 steps earlier on average than the EH-heavy policy $(600, 1200)$. All 12 parent-level contrasts are positive (exact one-sided sign-flip $p \approx 0.0002$). Both policies start at total pressure 1800, so this contrast isolates allocation rather than an increase in total initial contraction. It compares fixed coefficients selected at the fork, not policies with equal cumulative exposure. All source-swap arms are observed by the cutoff (Appendix M).

Both groups contribute, with a larger GC response. Increasing GC pressure from 600 to 1200 advances mean arrival by roughly 2,400–3,000 steps at each EH level. EH also accelerates arrival, with a smaller response. GC high-minus-low contrasts agree in 12/12 parents; EH contrasts agree in 12/12 at $q_{\mathrm{GC}} \in \{900, 1200\}$ and 10/12 at $q_{\mathrm{GC}} = 600$ (Figure 2).

In the specified pressure units, the additive estimate is $w_H = (0.1199, 0.0196)$, with 95% intervals $[0.1049, 0.1400]$ for GC and $[0.0138, 0.0249]$ for EH. This is a unit-dependent response comparison: both groups contribute, and GC has the larger response in these units.

Sensitivity to dose units and group composition. The bounded-score ratio is approximately 6:1 in pressure units, not an intrinsic efficiency ratio. Its magnitude changes substantially in decay-vector, fork-local Adam, and per-parameter units.

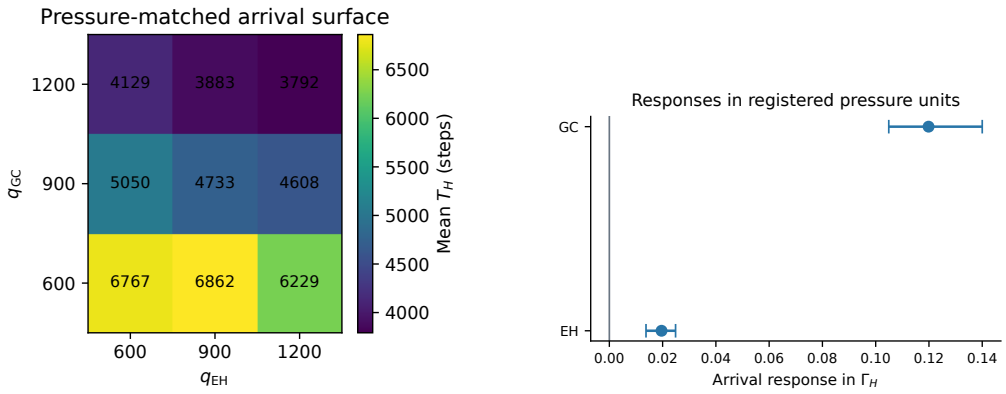


Figure 2: Primary pressure factorial. Left: calendar-censored arrival Y_H^{cal} at $C_H = 12{,}000$. Right: additive bounded-score responses with 95% parent-bootstrap intervals. Both group responses are positive; GC is larger in the specified units.

Across these coordinates, GC exceeds EH in 11/12–12/12 parents (Appendix G). These comparisons re-express executed policies; they do not test new interventions matched in alternative units.

LayerNorm contributes 29.4% of fork-time GC squared norm. Removing it from the dose coordinate changes the ratio from 5.8 to 8.3. In a separate experiment with LayerNorm and head decay disabled, core and embedding decay each advance arrival ($+1{,}727$ and $+969$ steps, $p = 0.019$ each); their difference is $+758$ steps (CI $[-625, 2088]$, $p = 0.194$; Appendix I). Section 9 discusses the interpretation of these bundled interventions.

5 From Completion Gaps to Delayed Arrival

Does the GC advantage mean earlier generalization, or failure to generalize under the alternative policy? We first compare a common budget across checkpoint states, then extend the delayed arms. A fresh 16-parent cohort forks at 200 steps before baseline memorization onset (pre-M), at onset (M-onset), and 450 steps after onset (post-M). The forks are aligned retrospectively to each baseline run's memorization event. Four cells at $q_{\mathrm{GC}}, q_{\mathrm{EH}} \in \{600, 1200\}$ yield 192 arms. At each stage, Equation 1 sets fixed continuation coefficients from the stage-specific norms. Thus the cross-stage comparison changes both checkpoint state and raw decay coefficients; it does not isolate a pure state effect under identical coefficients (Appendix F).

5.1 At the same budget, early forks show a completion gap

At $L = 8{,}000$, high-GC arms arrive in $31/32$ pre-M and $32/32$ M-onset continuations; low-GC arms arrive in $0/32$ at both stages. Pooling early contexts gives $63/64$ versus $0/64$, a descriptive risk difference of 0.984. The parent-level GC completion contrasts are 0.969 and 1.0, respectively, with all 16 parents positive ($p = 2^{-16}$ for each stage). Inference uses the 16 independent parents, with continuations paired within each parent. All 64 pre-M arms subsequently memorize, so the completion gap is not a failure to fit the training set.

Post-M, both GC conditions reach arrival in $32/32$ arms by the same 8k budget, but high GC advances arrival by 2,719 steps. The paired bounded-score contrast is 0.340 (16/16 positive, $p = 1.53 \times 10^{-5}$; bootstrap lower bound 0.320). EH responses remain positive. The ordering persists while the observed outcome changes from a completion gap to a timing difference.

At 4k, even post-M shows $31/32$ versus $0/32$ completion; by 6k, $13/32$ low-GC post-M arms arrive. The completion gap therefore changes with the observation budget. The stage-specific raw-coefficient remapping is also explicit: under three alternative fork-local units, post-M GC contrast sizes are 1.19–1.73 times the pre-M sizes, rather than attenuated (Appendix G).

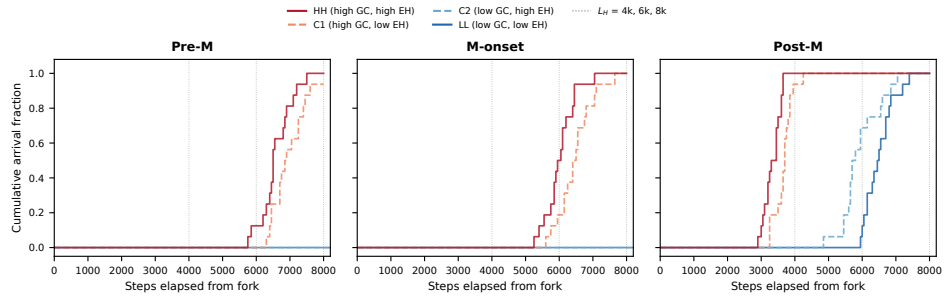


Figure 3: Arrival incidence by fork stage (32 arms per curve: 16 parents $\times$ two EH levels). At 8k, early low-GC arms remain unobserved, whereas both post-M conditions have arrived. Curves show arm-level incidence; inference is paired at the parent level.

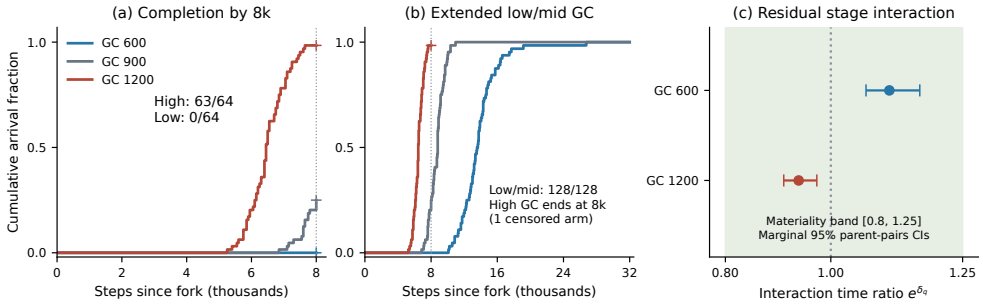


Figure 4: Direct arrival observations and secondary model inference. (a) Early 8k completion gap. (b) All 128 extended low-/mid-GC arms arrive by 32k; high GC stops at its actual 8k cap (one right-censored arm). (c) Corrective AFT reanalysis: interaction time ratios e^{δ_q} relative to the middle dose, with marginal parent-pairs 95% intervals. The comparison log band maps to $[0.8, 1.25]$. Curves are descriptive arm-level summaries.

5.2 Longer observation resolves the delayed arms

A registered midpoint extension first added $q_{\mathrm{GC}} = 900$ at all three stages and both EH levels (96 arms), giving 288 distinct parent–stage–dose conditions. A separate long-horizon extension then reforked the early $q_{\mathrm{GC}} \in \{600, 900\}$ conditions at both EH levels to 32,000 steps (128 arms). All 128 arrived, and every first-8k behavioral prefix exactly reproduced its historical counterpart. The extended low-/mid-GC runs therefore exhibit delayed arrival. Early high-GC arms retained their 8k cap, including one censored arm. Replacing the corresponding short records with their extended records, rather than counting both, gives 288 records: 287 observed and one censored (Appendix L).

5.3 Timing differences and deadlines

A large completion gap need not imply a large change in the timing response: similar accelerations can cross a deadline from different baselines. An exploratory log-normal accelerated failure-time (AFT) analysis illustrates this distinction (Figure 4c). A common GC timing profile, retaining parent$\times$stage baselines and stage-specific EH slopes, already describes a large early completion separation at 8k. Allowing early-stage GC interactions improves fit, but the two reference-dose interactions correspond to approximately $+11\%$ and -6% shifts in time. This is a model-based description, not a causal decomposition or proof of stage invariance.

The null calibration was repaired after inspecting outcomes, and the common-horizon reconstruction is post hoc. Neither is needed to establish arrival of the extended arms. Appendix L reports the conditional test, parent-pairs intervals, correction record, and full model comparison (Table 8).

6 A FIXED ALLOCATION SWAP ON A SECOND TASK

Does the allocation effect extend beyond modular addition? On S_5 permutation composition (120 classes), the same architecture and a separate 12-parent step-800 fork test the equal-total swap in Equation 3. The Z97 source contrast was fixed before target outcomes. At common calendar cutoff $C_H = 12{,}000$, GC-heavy advances arrival by $1{,}063$ steps (10 positive, one negative, one zero; primary exact mean sign-flip $p = 0.005$; parent-bootstrap CI $[450, 1621]$). At 32k, both swap arms are observed for every parent: the advance is $1{,}688$ steps, 11/12 positive, $p = 0.002$ (Appendix M). Thus the allocation advantage extends to a second task and remains after the swap arms are fully observed.

The 12k estimate is a bounded-time contrast, not an uncensored mean event-time difference. A post-hoc confirmation-time sensitivity retains the positive direction at both cutoffs (Appendix M); it does not replace the primary endpoint. This tests transport of one fixed allocation contrast, not the full response surface or its magnitude. The high/high cell has the earliest mean arrival in the primary Z97 grid: allocation is a distinct control, not a substitute for total dose.

Coefficient-local replications at width 96 (18/18 positive GC contrasts) and modular multiplication provide additional directional evidence under different units (Appendices O–P). From initialization, the dose-averaged GC-only versus EH-only contrast is positive in 16/16 seeds, but high-dose completion is sparse and stable-onset ordering descriptively reverses (Appendix U). The evidence therefore does not support an unconditional GC-over-EH ranking.

7 TRAINING HISTORY: ARRIVAL AND IMMEDIATE RESPONSE

The preceding experiments change decay after a shared checkpoint. We now hold the late policy fixed and vary inherited parameter and optimizer sources, asking whether earlier allocation history affects future arrival. An independent one-step assay then characterizes how these source states respond locally to decay.

7.1 CHANGING PARAMETER SOURCE CHANGES FUTURE ARRIVAL

Under a common continuation policy, switching the parameter source from an EH-heavy to a GC-heavy history advances arrival by 810 steps on average. To measure this history effect, a fresh 32-parent cohort constructs two early histories and independently assembles parameter source θ and complete Adam-history source O before the common late policy. G/E label the GC-heavy/EH-heavy source histories, not parameter subgroups being spliced. Each parent supplies

$$Y_{GG} = Y[\theta_G, O_G], \quad Y_{GE} = Y[\theta_G, O_E], \quad Y_{EG} = Y[\theta_E, O_G], \quad Y_{EE} = Y[\theta_E, O_E].$$

The four arms share continuation controls, including RNG and batch stream. Their mean arrival times are $4{,}358$, $4{,}484$, $5{,}253$, and $5{,}209$ steps, respectively. The parameter-source contrast $\tau_{\theta,p} = (Y_{GG,p} - Y_{EG,p} + Y_{GE,p} - Y_{EE,p})/2$ averages -810 steps (31/32 negative; two-sided exact sign-flip $p = 9.31 \times 10^{-10}$). Simple-effect means agree under both optimizer sources: $Y_{GG} - Y_{EG} = -895$ (32/32) and $Y_{GE} - Y_{EE} = -725$ (31/32).

The optimizer-source main-effect estimate is $\bar{\tau}_O = -41$ steps ($p = 0.365$). Interaction $I = Y_{GG} - Y_{GE} - Y_{EG} + Y_{EE}$ averages -170 (Holm $p = 0.042$), below the 500-step materiality criterion and short of the 24/32 concordance criterion (21/32).

The parameter-source effect has the same direction under both optimizer sources. However, compatibility calibration did not produce an admissible threshold, so the operation does not isolate history storage independently of parameter–optimizer mismatch (Appendix N).

7.2 INDEPENDENT GEOMETRY CHARACTERIZES THE IMMEDIATE RESPONSE

Why might the same kind of decay intervention produce different immediate output changes after different histories? Two factors can differ: the parameter change written by decay, and the network's local mapping from parameter changes to outputs. A separate cohort of 24 eligible parents measures these factors in a one-step assay. This is not a mediation analysis of the 810-step arrival effect, and it shares no parents with that experiment.

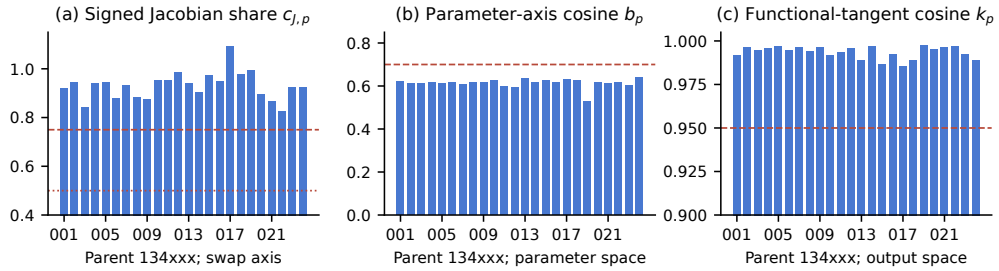


Figure 5: Independent fresh geometry cohort ($n = 24$). Left: signed Jacobian contribution c_J; values above one indicate cancellation. Center: worst-source absolute parameter-axis cosine b_p. Right: worst-cell functional-tangent cosine k_p. H1 and H2 pass their specified thresholds; all 24 parents satisfy the joint H2 criterion.

A source-normalized decay pulse defines a parameter-space write vector B and output tangent $K = J_Z B$, where J_Z is the Jacobian of the full centered-logit readout at the assay's one-step evaluation point. Subscripts G/E below denote GC-heavy/EH-heavy *source histories*, not parameter groups. Here J_G, J_E average the Jacobians over the two optimizer contexts for each parameter source; B is shared across those contexts. With $\bar{J} = (J_G + J_E)/2$ and $\bar{B} = (B_G + B_E)/2$, the symmetric bilinear expansion is exact:

$$T^\theta = J_G B_G - J_E B_E = \underbrace{(J_G - J_E)\bar{B}}_{D_J} + \underbrace{\bar{J}(B_G - B_E)}_{D_B}, \qquad c_J = \frac{\langle D_J, T^\theta \rangle}{\|T^\theta\|^2}. \tag{4}$$

Defining c_B analogously gives $c_J + c_B = 1$. The empirical question is which term dominates this signed projection.

H1: source-contrast decomposition. On the swap axis, median $c_J = 0.938$, with 24/24 parents above one half. This signed projection share describes the one-step source contrast. Values above one indicate cancellation by the other term.

H2: distinct parameter directions, near-parallel functional tangents. We compare an allocation direction (increasing GC while decreasing EH) with a total-pressure direction (increasing both). Their tangents are $K^{\text{swap}} = K_{\text{GC}} - K_{\text{EH}}$ and $K^{\text{total}} = K_{\text{GC}} + K_{\text{EH}}$. This asks whether two different parameter interventions remain distinguishable by their immediate output directions. The median worst-source absolute parameter-axis cosine is $b_p = 0.615$; the median worst-cell functional-tangent cosine is $k_p = 0.995$. All 24 parents pass the joint criterion: distinct parameter directions map to nearly parallel output tangents (Figure 5). Appendix Q gives the definitions, criteria, and per-parent results. Near-parallel output directions need not have equal magnitudes or arrival effects. They can also arise when one group-response component dominates both combinations, so alignment alone does not identify a low-dimensional mechanism.

Together, H1 and H2 characterize how inherited parameter sources shape the immediate response to decay.

8 Discussion

Where decay is applied changes generalization timing. Uniform decay moves both groups together; the equal-total swap shows that its allocation is a separate experimental control. The 2,100-step Z97 advance and source-fixed S_5 test make this point without relying on a coordinate-invariant response ratio. Both groups help within the factorial, so the result is not a claim that EH is irrelevant or that GC decay is necessary.

Observation time shapes the apparent generalization gap. The early 8k completion gap and eventual arrival of all extended runs describe the same training trajectories at different horizons. This establishes state- and horizon-conditioned outcomes, not a large intrinsic stage interaction under

an identical intervention operator. A deadline-specific success rate remains useful, but it answers a different question from eventual arrival. Grokking intervention studies should report the event rule, observation horizon, and paired timing contrast alongside completion fractions; when feasible, following censored alternatives distinguishes delay from unresolved non-arrival.

From immediate response to arrival. The geometry experiment shows how inherited parameter sources shape the immediate response to decay, with the Jacobian term dominating the signed source contrast. The mechanistic question is how this response evolves under continued optimization and relates to eventual arrival (Appendix R).

9 LIMITATIONS

Dose and group interpretation. Pressure matches initial decay-only contraction, not functional perturbation, total optimizer updates, or cumulative exposure (norm-path audit: Appendix M). Coordinate sensitivities re-express executed policies, rather than testing policies matched in alternative units. They establish neither universal GC dominance nor necessity of either group. GC bundles attention, MLP, and LayerNorm; its effect cannot be attributed to one component or function. Removing LayerNorm from a coordinate does not disable its decay; the separate core–embedding contrast remains unresolved. Cross-stage comparisons also change raw coefficients, as noted in Section 5.

Observation and inference. The 32k finding covers only the 128 extended low-/mid-GC arms; one unextended high-GC arm remains censored. Cohort-specific timestamps can change bounded contrasts (Appendix C). AFT inference is exploratory, with post-outcome calibration repair and post-hoc common-horizon reconstruction. Its test is model-conditional, not cluster-robust population inference; fitted counts are in-sample descriptions, not causal decomposition or independent validation. Marginal time-ratio intervals do not establish simultaneous equivalence. The direct arrival finding does not depend on this model. Parent-level pairing respects within-parent dependence, but 12–16-parent primary cohorts limit population extrapolation.

History and mechanism. The 810-step transplant effect is specific to the source-changing operation: failed compatibility calibration leaves parameter–optimizer mismatch unresolved, not proven (Appendix N). It does not isolate history storage independently of compatibility. Nonsignificant optimizer effects and subthreshold interactions do not establish zero effects. In the one-step assay, c_J is a signed projection share, not explained variance or a fraction of the arrival effect. Swap/total alignment implies neither GC/EH alignment nor globally low-dimensional dynamics. No geometry-to-arrival mechanism or universal GC/EH persistence ordering is established; both require tracking the evolving parameter–optimizer state.

Additional predictive and mechanism audits do not identify a sufficient state summary or a validated mediator of arrival (Appendices H, S, and T). The additive summaries do not test GC$\times$EH nonadditivity. Circuit, norm, and optimizer explanations remain open.

Applicability. The evidence concerns small two-layer Transformers, selected checkpoint states, and the tested algebraic tasks and optimizer settings. Event-aligned forks are retrospective rather than online policies. The S_5 experiment tests a fixed allocation swap, not transport of the whole response surface or effect magnitude; whole-training high-dose results also show a descriptive reversal of stable-onset ordering. Comparisons with tuned uniform AdamW and no-decay baselines are needed to establish practical training benefits.

10 CONCLUSION

At the same total initial decay pressure, shifting allocation toward GC advances grokking in paired modular-addition continuations and in a source-fixed S_5 test. Extending the slower early-fork runs shows that their striking short-budget completion gap is a timing gap, not permanent failure. Parameter-source and single-step assays provide complementary state-dependent evidence, not a mechanism for arrival. The central lesson is to distinguish where decay is applied, when intervention begins, and how long generalization is observed.

Reproducibility Statement. The anonymous analysis package is provided as `artifact.zip`. Its executable scope is analysis-level reproduction: behavioral endpoints and primary statistics are recomputed from released evaluation trajectories; transplant contrasts use released endpoint records; local-geometry statistics and gates use the frozen decision table. Training and checkpoint-level Jacobian/tangent measurements are not rerun by the package. The claim and figure maps distinguish recomputation from verification of retained summaries and document differences from manuscript figure layouts.

All primary cohorts use frozen protocols and immutable pre-execution bindings (seed pools, pressure levels, eligibility criteria, endpoint definitions, and gate criteria frozen before any continuation arm launches). Every continuation arm in the non-transplant cohorts is checkpoint-cloned: model weights, Adam optimizer moments, batch generator state, and RNG state are restored from a common source checkpoint. The transplant cohort instead prospectively constructs hybrid states by independently crossing parameter source and optimizer-history source under its frozen protocol, while holding the common late policy and registered within-parent continuation controls fixed. Checkpoint digests, split digests, and expected batch-stream digests are recorded. The extended-horizon refork cohort (128 arms) was executed under a separate frozen registry and protocol; prefix-replay exactness was verified for all 128 arms (Appendix L). The S_5 cross-task cohort (48 arms) was executed under a separate frozen registry (protocol v1.2.1); the Z97 source-derived contrast vector was hash-bound before any S_5 target outcome was generated (Appendix M). The transplant factorial cohort (32 fresh parents, 128 arms, seeds 133001–133032) was executed under a separate frozen protocol; an aborted short-run attempt was excluded (Appendix N). Behavioral evaluation trajectories recorded every 50 steps are released as `.npz` files; these are not model-weight checkpoints. Analysis scripts, sign-flip test implementations, frozen protocols, registries, and manifests are provided in the supplementary code. Sign-flip tests and parent-block bootstraps operate at the parent level; the AFT inference separates a conditional parametric null calibration from observed-parent pairs intervals (Appendix L).

The anonymous `artifact.zip` includes the endpoint-description audit under `audits/endpoint/`. This artifact-tree edition derives from the historical v566 audit and documents cohort-specific event timestamps, the S_5 protocol/implementation boundary, and post-hoc confirmation-time sensitivities.

The post-hoc unified-8k AFT comparison and sensitivity analysis have separate input/code hash bindings, 10,000 retained parent-pairs draws, and formal analysis and audit scripts (Appendix L); this analysis is not a prospectively registered training experiment.

The anonymous release is a single package with three components: (i) the core intervention artifact for the arrival and transplant results (Appendices A, B, and D); (ii) the nested analysis-level supplement for the extended stage and transport audits (Appendices S–U); and (iii) the embedded fresh-confirmation assets for the independent local-geometry results in Section 7.2 and Appendix Q, containing the frozen protocol with its three-stage freeze chain, the launch-authorization and audit records, measurement and decision code, the hash-bound decision table, compact results, and verification scripts. Large raw checkpoint and tangent objects omitted from the fresh-confirmation assets remain identity- and hash-bound in their manifests. Analysis scripts, sign-flip test implementations, frozen protocols, registries, and manifests are provided across these components.

AI-Use Statement. AI-assisted tools were used to refine hypotheses and conceptual framing; provide feedback on experimental design and statistical methodology; assist with analysis-code implementation; support literature search; critique claim calibration and result interpretation; and edit or format portions of the manuscript. All protocols, analyses, numerical results, and final scientific conclusions were independently checked and approved by the authors. No AI tool generated experimental data.

REFERENCES

H. V. AlquBoj, H. AlQuabeh, V. Bojkovic, M. Nwadike, and K. Inui. Mechanistic insights into grokking from the embedding layer. *arXiv preprint arXiv:2505.15624*, 2025.

E. Boursier, S. Pesme, and R.-A. Dragomir. A theoretical framework for grokking: Interpolation followed by Riemannian norm minimisation. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

A. Gromov. Grokking modular arithmetic. *arXiv preprint arXiv:2301.02679*, 2023.

D. He, S. Tu, A. Jaiswal, L. Shen, G. Yuan, S. Liu, and L. Yin. AlphaDecay: Module-wise weight decay for heavy-tailed balancing in LLMs. *arXiv preprint arXiv:2506.14562*, 2025.

T. Kim. Grokking on the weight-decay clock: A rate hierarchy from softly broken symmetries. *arXiv preprint arXiv:2607.23967*, 2026.

T. Kumar, B. Bordelon, S. J. Gershman, and C. Pehlevan. Grokking as the transition from lazy to rich training dynamics. In *International Conference on Learning Representations*, 2024.

Z. Liu, O. Kitouni, N. Nolte, E. J. Michaud, M. Tegmark, and M. Williams. Towards understanding grokking: An effective theory of representation learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

N. Nanda, L. Chan, T. Lieberum, J. Smith, and J. Steinhardt. Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations*, 2023.

A. Power, Y. Burda, H. Edwards, I. Babuschkin, and V. Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. In *Mathematical and Empirical Understanding of Foundation Models Workshop at NeurIPS*, 2022.

L. Prieto, M. Barsbey, P. A. M. Mediano, and T. Birdal. Grokking at the edge of numerical stability. In *International Conference on Learning Representations*, 2025.

D. Stander, Q. Yu, H. Fan, and S. Biderman. Grokking group multiplication with cosets. In *International Conference on Machine Learning*, 2024.

X. K. Truong, Q. H. Truong, D. T. Luu, and T. D. Phan. The norm-separation delay law of grokking: A first-principles theory of delayed generalization. *arXiv preprint arXiv:2603.13331*, 2026a.

X. K. Truong, H. V. Doan, D. T. Luu, and T. D. Phan. The weight norm sets the grokking timescale: A causal delay law. *arXiv preprint arXiv:2606.13753*, 2026b.

X. K. Truong. Cross-trajectory chimera interventions reveal dissociable roles of weight magnitude and direction in grokking. *arXiv preprint arXiv:2607.06628*, 2026c.

X. K. Truong. What does the weight norm control in grokking? Logit-scale mediation under cross-entropy. *arXiv preprint arXiv:2606.18465*, 2026d.

L. Verma. Weight decay regimes in grokking transformers: Cheap online diagnostics. *arXiv preprint arXiv:2605.20441*, 2026.

V. Varma, R. Shah, Z. Kenton, J. Kramár, and R. Kumar. Explaining grokking through circuit efficiency. *arXiv preprint arXiv:2309.02390*, 2023.

Y. You, I. Gitman, and B. Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.

Y. You, J. Li, S. Reddi, J. Hseu, S. Kumar, S. Bhojanapalli, X. Song, J. Demmel, K. Keutzer, and C.-J. Hsieh. Large batch optimization for deep learning: Training BERT in 76 minutes. In *International Conference on Learning Representations*, 2020.

Z. Zhong, Z. Liu, M. Tegmark, and J. Andreas. The clock and the pizza: Two stories in mechanistic explanation of neural networks. In *Advances in Neural Information Processing Systems*, volume 36, pages 27223–27250, 2023.

Z. Liu, E. J. Michaud, and M. Tegmark. Omnigrok: Grokking beyond algorithmic data. In *International Conference on Learning Representations*, 2023.

V. Thilak, E. Littwin, S. Zhai, O. Saremi, R. Paiss, and J. Susskind. The slingshot mechanism: An empirical study of adaptive optimizers and the grokking phenomenon. *arXiv preprint arXiv:2206.04817*, 2022.

A Primary Intervention Cohorts

Table 2: Primary intervention cohorts.

Cohort	Parents	Arms	Estimand
Pressure-matched factorial	12	108	Group-pressure response
Three-stage factorial	16	192	Stage-dependent accessibility/timing
Extended-horizon refork	16 (reused)	128	Censor-repair AFT interaction
Cross-task S_5 swap	12 fresh	48	Task transport of finite-secant direction
Parameter-state transplant	32 fresh	128	Operational parameter-source arrival effect
Fresh local geometry	24 fresh	96	Registered one-step H1/H2 properties

B Experimental Protocol

Task. The task is modular addition over $p = 97$: the input grid contains $p^2 = 9409$ ordered pairs (a, b) with target $(a + b) \bmod p$. For every seed we form a deterministic disjoint split of 30% training, 20% calibration (excluded from gradient training; not used in the primary analyses reported here), and 50% final-test pairs.

Architecture. The network embeds the two input tokens into width 128, applies two causal Transformer blocks with four attention heads and feed-forward width 512, takes the final-token hidden state after LayerNorm, and maps it to 97 logits with a linear head; dropout is zero.

Optimizer. Training uses AdamW with learning rate $\eta = 10^{-3}$, betas $(\beta_1, \beta_2) = (0.9, 0.999)$, batch size 512, and constant weight decay applied per parameter group. Phase-map runs are evaluated every 50 optimizer steps. For the checkpoint-fork cohorts, parents are trained from initialization under uniform $w_d = 1$; only the fork continuation changes treatment. Appendix U separately applies group-selective decay from initialization.

Parent-state fork. Each fork restores parameters, Adam moments, batch-generator state, and random-number state before changing only the registered treatment. All primary pressure-matched, three-stage, midpoint-extension, and extended-horizon cohorts passed their registered source-state, optimizer-state, and stream-consistency audits (zero model and optimizer discrepancy and matching RNG digests). The fork point is the first eligible step-800 memorization checkpoint; formal memorization confirmation is evaluated separately by the three-checkpoint rule. Fork continuations use 50-step evaluation spacing.

Parent eligibility gate. Candidate seeds must satisfy at step 800: train accuracy ≥ 0.99, test accuracy < 0.10, and cross-entropy relative drift ≤ 0.50 over a 3-checkpoint window (steps 700–900). Seed IDs are frozen before any continuation arm is executed.

Checkpoint cadence. Behavioral checkpoints are saved every 50 steps. Event times are recomputed deterministically from the saved NPZ trajectories; no parameter is refit on continuation data.

C Event and Censoring Definitions

Memorization (M). The first 3-checkpoint window at which train accuracy ≥ 0.99, test accuracy < 0.10, and train cross-entropy relative drift ≤ 0.50 hold simultaneously. The fork point is the first M-confirmed checkpoint at step 800.

Episode onset and confirmation. A qualifying H episode consists of five consecutive 50-step evaluations with train accuracy ≥ 0.99 and test accuracy ≥ 0.90. Let t_o be its first checkpoint and $t_c = t_o + 200$ its fifth (confirmation) checkpoint; absent an episode, both are infinite. Calendar cutoff $C = t_f + L$ corresponds to elapsed budget L. Where retained tables and analyses use T_H,

it denotes the timestamp of that cohort, not a universal confirmation time. Three implemented rules appear in the reported experiments:

$$Y_C^o = \min(t_o, C), \qquad I_C^o = \mathbf{1}\{t_o \le C\}, \tag{5}$$

$$Y_C^c = \min(t_c, C), \qquad I_C^c = \mathbf{1}\{t_c \le C\}, \tag{6}$$

$$Y_C^g = \begin{cases} t_o, & t_c \le C, \\ C, & t_c > C, \end{cases} \qquad I_C^g = \mathbf{1}\{t_c \le C\}. \tag{7}$$

Rule o uses later recorded evaluations to verify a window beginning by C. Rule c timestamps confirmation. Rule g timestamps onset but requires confirmation by C; it is a bounded operational score, not simply $\min(t_o, C)$. Subtracting t_f from any calendar score gives its elapsed version. This change of time origin preserves paired differences; changing endpoint rules need not.

Table 3: Endpoint conventions of the principal behavioral cohorts. Historical reported values retain the implemented rules; the confirmation-time checks in this appendix are post hoc.

Cohort	Scored timestamp	Observation rule
Primary pressure factorial	Onset (o)	Onset by C; confirmation checked in the longer record
Stage, midpoint, and refork	Confirmation (c)	Confirmation by the arm's cutoff
Parameter-source transplant	Onset (o)	Onset by C; confirmation within the execution look-ahead
S_5 transport	Gated onset (g)	Confirmation by C at both reported cutoffs

Endpoint-description correction and protocol boundary. Earlier manuscript versions described all primary arrival times as fifth-checkpoint confirmation times. That description did not match the primary factorial, transplant, and S_5 implementations. We correct the description while retaining their original outputs. The factorial's `events.py` explicitly tests onset against the cutoff. The S_5 parser and its source-equivalence adapter instead test confirmation against the cutoff and return onset for observed arms. Thus their equivalence self-test establishes agreement of those adapters, not equality of all cohort endpoint rules. No primary-factorial arm has $t_o \le 12000 < t_c$, and the bound source swap contains only observed arms, so the adapter distinction does not alter that source vector.

The S_5 protocol v1.2.1 names onset as the timestamp and writes $Y = \min(T_H, C)$, but also requires confirmation by 12k; its 32k execution-look-ahead wording additionally permits an onset-based reading at the boundary. These statements are not identical for $t_o \le C < t_c$. The released implementation uses rule g at both cutoffs. We state this ambiguity explicitly rather than claiming that the protocol language determines a unique boundary rule. All swap arms are confirmed by 32k, so that boundary ambiguity does not affect the reported long-horizon swap. Appendix M reports a separate confirmation-time sensitivity.

When timestamp changes cancel. For a pair observed under both rules, changing onset to confirmation adds 200 steps to both arms, leaving the difference unchanged. Two arms censored under both rules also retain a zero difference. One observed and one censored arm generally do not cancel, even without an onset/confirmation boundary case. For example, at parent 321 and $q_{\mathrm{EH}} = 900$, the low-GC score remains 12000 while the high-GC score changes from 4400 to 4600: the contrast is 7600 under rule o and 7400 under rule c. The primary equal-total source swap is unchanged, as are the fully observed transplant factorial contrasts. This does not imply invariance of all factorial contrasts, additive slopes, absolute arrival means, or censored outcomes. The three EH-specific GC mean advances under the retained factorial endpoint are 2637.5, 2979.2, and 2437.5 steps.

Reproducible scope of this correction. The audit script `paper_v566_endpoint_audit_v1/code/audit_endpoints.py` reads the 108 primary-factorial metadata records, the saved S_5 per-parent endpoint records, and the transplant factorial table. It verifies retained scores, computes the explicitly post-hoc confirmation sensitivities, and records input/source-code hashes in `results/endpoint_audit.json`. It does not retrain models, replay raw trajectories, authenticate historical preregistration hashes, or change original analysis files. Supporting cohorts retain their own stated endpoint definitions.

Sign-count tests and mean-based sign flips. We distinguish the exact binomial sign-count test from the exact mean-based sign-flip test. For the sign-count test, zero paired differences are removed only for the test, $n_{\text{nonzero}} = n_{\text{pos}} + n_{\text{neg}}$, and the preregistered one-sided upper-tail p-value is

$$p^{+}_{\text{sign}} = \sum_{k=n_{\text{pos}}}^{n_{\text{nonzero}}} \binom{n_{\text{nonzero}}}{k} 2^{-n_{\text{nonzero}}}.$$

The corresponding two-sided sign-count value is the doubled smaller of the upper and lower binomial tails, capped at one. This sign-count test uses only the signs and does not use the magnitudes of the paired differences.

For a mean-based exact sign-flip test, the parent-level contrast vector $D = (D_1, \ldots, D_n)$ is retained as observed, including any $D_i = 0$ values. With $\bar{D}_{\text{obs}} = n^{-1} \sum_i D_i$ and the registered inclusive $\epsilon = 10^{-12}$, the one-sided upper-tail value is

$$p^{+}_{\text{SF}} = 2^{-n} \# \left\{ s \in \{-1, +1\}^n : \frac{1}{n} \sum_i s_i D_i \geq \bar{D}_{\text{obs}} - \epsilon \right\},$$

and the two-sided value is

$$p^{\pm}_{\text{SF}} = 2^{-n} \# \left\{ s \in \{-1, +1\}^n : \left| \frac{1}{n} \sum_i s_i D_i \right| \geq |\bar{D}_{\text{obs}}| - \epsilon \right\}.$$

Thus zeros remain in the denominator 2^n, while the effect magnitudes are retained in the test statistic. For an all-positive directional contrast this reduces to $p = 2^{-n}$, but it is not generally equal to the binomial sign-count value when magnitudes differ or zeros are present.

For the S_5 transport analysis, the registered parent contrast is $D_i = Y_{H,i}(600, 1200) - Y_{H,i}(1200, 600)$. The primary test uses the first mean-based formula with $n = 12$ and retains one zero contrast: its exact value is $22/4096 = 0.00537109375$, reported in the main text as $p = 0.005$. The registered secondary sign-count test removes that zero and gives $12/2048 = 0.005859375$.

For the transplant factorial, with cells $A_p = Y_p[\theta_G, O_G]$, $B_p = Y_p[\theta_G, O_E]$, $C_p = Y_p[\theta_E, O_G]$, and $D_p = Y_p[\theta_E, O_E]$, the parent-level contrasts are

$$\begin{aligned} \tau_{\theta,p} &= \frac{A_p - C_p + B_p - D_p}{2} = 25 z_{\theta,p}, \\ \tau_{O,p} &= \frac{A_p - B_p + C_p - D_p}{2} = 25 z_{O,p}, \\ I_p &= A_p - B_p - C_p + D_p = 50 z_{I,p}. \end{aligned}$$

The primary, optimizer, and interaction p-values use the two-sided mean-based formula on their corresponding integer-scaled vectors; the primary dynamic- programming enumeration gives $4/2^{32} = 9.313225746 \times 10^{-10}$, reported as 9.3×10^{-10}. No reported result or decision threshold is changed by this distinction.

Finite-sample correction for harmonic-10. The separate ten-parent harmonic-anisotropy audit reports a conventional plus-one version of the same one-sided mean-based sign-flip tail. If r is the number of permutations at least as extreme as observed, its reported value is $(r + 1)/(2^n + 1)$, with the same inclusive ϵ rule and with zeros retained if present. This is a finite-sample reporting correction, not the uncorrected p^{+}_{SF} definition above. For the harmonic-10 vector all ten contrasts are positive, so $r = 1$ and the reported value is $2/1025 = 0.0019512195$, shown as $p = 0.002$ in Appendix T. All other mean-based sign-flip values in this paper use the uncorrected exact tail defined above.

D Allocation Cohort Provenance

A prior fixed-target allocation scan (19 parents, 95 arms; 3 blocks: seeds 9–14, 21–36, 101–112) motivated the pressure-matched experiment. Its protocol, solver coefficients, and per-seed event times are available in the frozen repository but are not part of the primary inferential chain.

E PRE-FORK NORM AND DECAY-IMPULSE AUDIT

The pressure-matched factorial fixes group pressures at the fork, but groups differ in raw norm. At $t_{\text{fork}} = 800$:

Group	$\|\theta_g\|_2$	$S_g = \|\theta_g\|_2^2$	Tensor count
GC (attn+MLP+LN)	≈ 31.5	≈ 989	26
EH (emb+head)	≈ 50.9	≈ 2595	4

EH has fewer parameter tensors but a larger fork-time norm ($S_{\text{EH}}/S_{\text{GC}} \approx 2.6$). At budget-solver doses the GC pressure exceeds EH by 70–128×. On the pressure-matched grid, the ratio $q_{\text{GC}}/q_{\text{EH}}$ is held at $\{0.5, 1, 2\}$ by design, removing the 70–128× asymmetry that would otherwise confound raw coefficient comparisons.

F FORK-STAGE CONFIRMATION PROTOCOL

Cohort and design. The fork-stage confirmation (Section 5) is a separate preregistered experiment. It uses 16 fresh parents (seeds 120001–120016) drawn from an ordered candidate pool of 32 seeds (120001–120032), with no overlap with any discovery, pressure-matched, or D0-audit cohort. All 16 eligible parents were locked before any continuation arm launched; no replacement was permitted after the lock.

Each parent is forked at three event-relative stages anchored to that parent's own baseline memorization onset T_M^{baseline}:

Stage	Label	Δ from T_M	Role
Pre-M	`baseline_pre_M`	-200	before onset
M-onset	`baseline_M_onset`	0	at onset
Post-M	`baseline_post_M`	$+450$	post-M context

At each stage the four registered cells form a 2×2 factorial: LL (q_{GC}=600, q_{EH}=600), C1 (1200, 600), C2 (600, 1200), HH (1200, 1200). The fork-local coefficients are $\lambda_g = q_g/S_g(t_{\text{fork}})$, computed once at the fork and held fixed for the entire continuation. The continuation horizon is $L_H = 8000$ steps.

Endpoints. Binary accessibility is $I_H = \mathbf{1}\{T_H^{\text{confirm}} \leq t_{\text{fork}} + 8000\}$, where T_H^{confirm} requires train accuracy ≥ 0.99 and test accuracy ≥ 0.90 for 5 consecutive checkpoints (200 steps). The elapsed timing score is $Y_H^{\text{el}} = \min(T_H^{\text{confirm}} - t_{\text{fork}}, L_H)$ and $\Gamma_H^{\text{FC}} = 1 - Y_H^{\text{el}}/L_H$ (see §3 for the Y_H^{el}/Y_H^{cal} distinction). In the stage experiment L_H=8,000; in the pressure-matched and S_5 experiments L_H=11,200 (fork at step 800, C_H=12,000). Paired contrasts D_i are invariant to subtracting the common fork time, not generally to switching onset and confirmation timestamps (Appendix C).

Gate hierarchy and statistical inference. The confirmation uses a three-gate hierarchical procedure:

- **Gate 1 (primary):** Early GC accessibility. Tests $G_I(\text{pre})$ and $G_I(\text{onset})$ via one-sided exact parent-level sign-flip tests at $\alpha = 0.025$ each.
- **Gate 2 (primary):** Early-specific stage modulation. Tests $M^{\text{pre}} = G_I(\text{pre}) - G_I(\text{post})$ and $M^{\text{onset}} = G_I(\text{onset}) - G_I(\text{post})$ with the same method and $\alpha = 0.025$.
- **Gate 3 (key secondary):** Post-M GC timing. Tests $G_{\Gamma_H}(\text{post})$ at $\alpha = 0.05$.

The hierarchical decision rule was satisfied: each registered contrast was positive in $16/16$ parents, so each corresponding sign-flip test yields $p = 2^{-16} = 1.53\times10^{-5}$. The percentile lower confidence bounds are positive for all gates: $G_I(\text{pre})$ LCB $= 0.906$, $G_I(\text{onset})$ LCB $= 1.0$, $G_{\Gamma_H}(\text{post})$ LCB $= 0.320$.

G Coordinate and Intervention-Unit Sensitivity

Registered coordinate and sensitivity coordinates. The registered pressure coordinate $q_g^{\mathrm{reg}} = \lambda_g \|\theta_g\|_2^2$ approximates the exact one-step Euclidean squared-norm contraction mass $m_g^{\mathrm{Euc}} \approx 2\eta q_g^{\mathrm{reg}}$. We audit two further Euclidean conventions ($q_g^{(\Delta)} = \lambda_g \|\theta_g\|_2$ and $q_g^{(\mathrm{avg})} = q_g/d_g$) and a frozen fork-local Adam metric ($q_g^{\mathrm{Adam}} = \lambda_g \sum_{j \in g} G_j \theta_j^2$). Table 4 reports the results. The pooled GC slope is larger under every listed coordinate, while the ratio ranges from approximately 2 to 92 and parent-wise dominance from 11/12 to 12/12.

Table 4: Arrival slope ratio under registered and sensitivity coordinates (12-parent pressure-matched factorial, 107 arrived arms).

Coordinate	Ratio [95% CI]	GC stronger
Registered q^{reg}	5.83 [4.39, 8.60]	12/12
Exact Euclidean m^{Euc}	5.83 [4.39, 8.60]	12/12
Decay-update $q^{(\Delta)}$	3.69 [2.79, 5.43]	12/12
Per-parameter $q^{(\mathrm{avg})}$	91.88 [69.19, 135.45]	12/12
Fork-local Adam q^{Adam}	1.99 [1.52, 2.91]	11/12

The per-parameter coordinate is a mechanical re-expression using runtime trainable-parameter counts (d_{GC}=396,800, d_{EH}=25,185); it is not an independent mechanistic effect size.

What pressure matching removes and does not remove. The pressure grid removes the most direct norm-scale objection to raw coefficient comparisons: each axis is defined through the same first-order squared-norm contraction proxy. Equal q_g does not imply equal decay-vector magnitude, optimizer-relative local norm, curvature, or downstream response. GC and EH also differ in tensor structure and functional role. We therefore treat q_g as an *ex-ante matching coordinate* for first-order Euclidean contraction mass, not as a unique physical or mechanistic reaction coordinate.

Cross-stage intervention-unit sensitivity. Because the same registered pressure maps to different raw decay coefficients at event-relative forks, we retrospectively re-expressed the already executed GC contrast of the three-stage factorial (Section 5; 16 parents $\times$ 3 stages) in three additional fork-local units: the raw AdamW coefficient $\lambda_g = q_g/S_g(t_{\mathrm{fork}})$, the Euclidean decay-vector magnitude $\lambda_g\|\theta_g\|_2$, and the frozen Adam-relative metric q_g^{Adam} defined above. Relative to pre-M, the post-M effective GC contrast was $1.42\times$, $1.19\times$, and $1.73\times$ respectively (parent-bootstrap 95% CIs $[1.40, 1.45]$, $[1.18, 1.20]$, $[1.62, 1.82]$); it is exactly $1.00\times$ in registered pressure by construction. The amplification is uniform across parents ($16/16$ with ratio > 1 under every alternative unit; minima 1.31, 1.14, 1.17), and replacing the first-order mass $2\eta q$ by the exact finite-step contraction $[1-(1-\eta\lambda_g)^2]\,S_g$ leaves all ratios at 1.00 up to $(\eta q)^2/S$ corrections. Thus the change from finite-budget early completion separation to postM timing modulation cannot be attributed to attenuation of the administered GC contrast under any audited intervention unit: postM forks deliver, if anything, a larger effective contrast, because S_{GC} shrinks after memorization so equal q maps to larger raw λ. This is a descriptive re-expression of frozen interventions, not a metric-free or mechanistic claim.

H Full Baseline Comparison: Scalar-Exposure and Realized-Dose Models

Interpretation. The total-exposure scalar baselines (RMSE $\geq$ 1379) predict substantially worse than the group-resolved model (RMSE 717), confirming that total-exposure summarization discards group-specific information. However, the nominal-dose indicator model (α categorical, RMSE 619) is nominally better than the group-resolved model, though the paired parent-level sign-flip test does not reach significance (p=0.46; $12/19$ folds favorable; 95% bootstrap CI of the mean RMSE difference: $[-62, +277]$). This means the group-resolved model does not provide a demonstrated *incremental* predictive advantage over a dose-only indicator. The scientific identification of GC-dominant arrival therefore rests on the pressure-matched factorial contrasts (Section 4), which identify the directional response by construction, not on exposure-based predictability comparisons. The

Table 5: Nested leave-one-parent-out (LOSO) RMSE for all baseline models (19 parent seeds, 92 complete-case arms). The group-resolved model uses realized group exposures ($B_{\mathrm{GC}}, B_{\mathrm{EH}}$); total-exposure baselines use B_{total} alone; dose-indicator models use the nominal decay coefficient α_{EH} without computing realized exposures.

Model	Predictor(s)	LOSO RMSE
Total-exposure scalar baselines		
Linear total (1 df)	B_{total}	1618
Quadratic total (2 df)	$B_{\mathrm{total}},\ B_{\mathrm{total}}^2$	1474
Natural spline, 3 knots	B_{total}	1379
Realized-dose indicator baselines		
α categorical (4 df)	$\mathbf{1}(\alpha_{\mathrm{EH}} \in \{0.01, \ldots, 0.10\})$	619
α linear (1 df)	α_{EH}	668
α + total (2 df)	$\alpha_{\mathrm{EH}},\ B_{\mathrm{total}}$	691
Total + fraction (2 df)	$B_{\mathrm{total}},\ B_{\mathrm{EH}}/B_{\mathrm{total}}$	732
Group-resolved (2 df)	$B_{\mathrm{GC}},\ B_{\mathrm{EH}}$	717

dose-indicator models are prognostic baselines, not causal contrasts: they summarize which coefficient levels correlate with faster arrival but do not hold matched pressure on one group while varying the other.

Joint arrival/maintenance rank-one audit. The retained audit uses the 108-arm, 12-parent primary factorial, with 107 observed arrivals and 107 available maintenance outcomes. It jointly fits standardized arrival score Γ_H and maintenance score $\Gamma_M = 1 - \text{dip rate}$ (inclusive $[T_H, T_H + 3000]$ checkpoint window), with parent intercepts and additive GC/EH slopes. The combined fit contains 215 outcome records. Constraining the two response vectors to a shared rank-one direction changes SSE from 102.64235 to 102.91538, an increase of 0.27303 (0.266%). Rank-one minus unrestricted AIC is -1.42886 and BIC is -4.79950; both criteria favor the restricted fit. These retained model comparisons do not establish rejection of rank one, and are not a calibrated rank-null p-value. Source: `iclr_penalty_matched_factorial_v1/results/maintenance_audit_v1.json`.

The rank restriction concerns shared directions across two additive outcome responses. It does not test an $x_{\mathrm{GC}} x_{\mathrm{EH}}$ interaction or nonlinear response curvature. The observed nine-cell grid and fixed-EH contrasts support finite group comparisons, but this audit does not establish an additive global response surface or a scalar sufficient training state.

I Subgroup Composition Sensitivity

Design. This factorial decomposes the registered GC group into attention/MLP-core (with LayerNorm decay disabled) and the EH group into embedding (with head decay disabled). It uses 12 frozen parents (311, 312, 313, 314, 316, 317, 318, 320, 321, 322, 324, 326), pressure levels $q_{\mathrm{core}}, q_{\mathrm{emb}} \in \{600, 890\}$, fork step 800, and horizon 18,000. All 48 arms completed; 32/48 reached arrival under this factorial's endpoint at the 12,000-step cutoff. The two experiments measure different directional derivatives of the same response surface: the original GC direction is $(1, 1, 0, 0)$ in (core, LN, embedding, head) space, while the new core direction is $(1, 0, 0, 0)$.

Summary. Both attention/MLP-core and embedding carry substantial positive arrival effects ($p = 0.019$ each). The core effect is larger in point estimate (ratio ≈ 1.8), but the core-minus-embedding contrast is not statistically resolved ($p = 0.194$; CI includes zero). Per-parent heterogeneity is substantial: four parents show embedding advantage (gap < -1000), six show a large core advantage (gap $> +1000$). The standardized core estimate ($\sim$ 1,787 steps at the 300-step pressure span) is close to the original full-GC effect ($\sim$ 1,923 steps), but the two experiments occupy different control subspaces, so this proximity does not identify the incremental contribution of LayerNorm decay.

Table 6: Composition-sensitivity factorial: per-parent arrival contrasts (steps). Δ_{core}: attention/MLP-core high−low. Δ_{emb}: embedding high−low. Positive Δ = accelerated arrival.

Parent	Δ_{core}	Δ_{emb}	Gap
311	$+950$	$+2000$	-1050
312	$+3075$	$+1525$	$+1550$
313	$+1800$	$+900$	$+900$
314	$+4175$	$+325$	$+3850$
316	-725	$+2075$	-2800
317	$+3650$	$+150$	$+3500$
318	$+2225$	$+975$	$+1250$
320	$+175$	$+2475$	-2300
321	$+325$	-325	$+650$
322	$+3150$	-600	$+3750$
324	$+3025$	$+175$	$+2850$
326	-1100	$+1950$	-3050
Mean	$+1727$	$+969$	$+758$
95% CI	$[760, 2667]$	$[410, 1521]$	$[-625, 2088]$
Sign	$10/2$	$10/2$	$8/4$
p (1-sided)	0.019	0.019	0.194

J CLAIM CALIBRATION TABLE

Table 7 summarizes each claim's status, supporting evidence, and boundary conditions.

Table 7: Final claim calibration: arrival, state-source, transport, and local-response claims.

Claim	Status	Boundary
Arrival and stage (Sections 4–5.3)		
GC-larger arrival response	Supported	Both slopes positive; audited-unit ordering, coordinate-dependent ratio
EH arrival effect nonzero	Supported	w_{EH}=0.020; 95% CI $[0.014, 0.025]$
GC control ordering across stages	Supported	16/16 parents at every contrast
8k completion contrast	**Primary**	$63/64$ vs $0/64$; risk diff 0.984; p=2^{-16}; horizon outcome, not permanent path separation
Extended low-/mid-GC deficit not permanent	Confirmed	128/128 extended low-/mid-GC arms arrived; one unextended high-GC arm remains censored
Common three-dose profile	Rejected	LR= 53.03; $p_{\mathrm{MC}} = 1/10001$; 288-record AFT; model-conditional null calibration
Common linear log-time slope	Rejected	LR= 42.54
Material cross-stage interaction	Not established	Marginal intervals within log band $\pm\log(1.25)$; no simultaneous equivalence claim
Stage-invariance (all forms)	Not fully rejected	Two parametric forms rejected; flexible models not tested
Intrinsic 6:1 sensitivity	Not established	Ratio is metric-dependent (≈2–92)
History/state-source audit (Section 7)		
Parameter-source future-arrival effect	Supported	$\bar{\tau}_\theta$=-810, $31/32$, p=9.3×10^{-10}
Optimizer-source main effect	Not established	$\bar{\tau}_O$=-41, p=0.365
Material parameter×optimizer interaction	Not established	$\lvert I\rvert$=170 $<$ 500; $21/32 < 24/32$
Compact sufficient fork state	Not established	No closure test; registered candidate-realization screening found no successful compact predictor (App. S)
State dimensionality	Not identified	Scalar progress remains compatible
Cross-task transport (Section 6)		
GC-over-EH control ordering Z97→S_5	Supported	Mean $+1{,}063$ steps, p=0.005; $11/12$ at 32k
Full response surface transports	Not established	Only diagonal swap tested
GC-free EH-only entry on S_5	Not observed	GC-free high-EH substitute not observed
Local response geometry (Section 7.2; fresh confirmatory cohort)		
Local-Jacobian term dominates one-step parameter-source contrast	Confirmed	c_J=0.938 median, $24/24$>0.5; swap axis only
Parameter axes non-collinear, functional tangents near-parallel	Confirmed	b_p=0.615, k_p=0.995, joint $24/24$; no mediation/persistence/state-dimension claim
Whole-training policy transport (Appendix U)		
Dose-averaged ordering from initialization	Supported	$\bar{D}_H$=0.141, $16/16$, p=1.53×10^{-5}
Dose-independent policy dominance	Not established	Low dose strong ($15/16$ vs $0/16$); high dose weak, stable-onset reversed
Practical superiority over uniform AdamW	Not tested	No uniform/no-decay benchmark registered

K REGISTERED MIDPOINT EXTENSION

This appendix reports a prospectively registered dose-resolution extension of the three-stage confirmation. The extension adds $q_{GC} = 900$—the arithmetic midpoint of the registered interval $[600, 1200]$—to the same 16 parents and three fork stages, yielding $16 \times 3 \times 2 = 96$ new continuation arms. The sole scientific question is whether midpoint curvature relative to linear interpolation differs between the two early fork contexts (pre-M, M-onset) and the post-memorization context.

For each parent p, stage s, and EH level e, the interpolation contrast is $D_{p,s,e} = \Gamma_H(900, e)_{p,s} - \frac{\Gamma_H(600,e)_{p,s} + \Gamma_H(1200,e)_{p,s}}{2}$. The sole pre-registered inferential test is the stage-curvature contrast $C_p^{\text{stage}} = D_p^{\text{early}} - D_p^{\text{post}}$. We apply a two-sided exact parent-level sign-flip test to the 16 values C_p^{stage} at $\alpha = 0.05$, enumerating all $2^{16} = 65{,}536$ sign assignments.

Results. D^{early} is negative in $16/16$ parents (mean -0.080); D^{post} is positive in $15/16$ parents (mean $+0.057$). The contrast C^{stage} is negative in $15/16$ parents (mean -0.137); the registered sign-flip test gives $T = 2.187$, $p = 6.10 \times 10^{-5}$. The extended-horizon reforks and censor-aware AFT analysis (Appendix L) directly adjudicate the censoring-versus-dynamics ambiguity raised here.

Recovery audit. All 96 arms trained successfully (768,096 safety checks passed). The manifest collection step crashed due to a path-handling bug; recovery was performed without retraining, reading on-disk trajectories and recomputing all endpoints from the original arrays. The recovered endpoints and primary statistic are deterministic functions of the retained trajectories.

L EXTENDED-HORIZON CENSORING-REPAIR REFORK

Motivation and design. The bounded-score curvature contrast C^{stage} (Section 5) measures the observed score shape at $L_H = 8{,}000$. Because early-stage low-GC arms were heavily censored ($0/32$ arrival), the observed curvature could arise from stage-dependent censoring rather than from a genuinely stage-dependent event-time distribution. To distinguish these, we prospectively reforked the same 16-parent cohort's pre-M and M-onset arms at $q_{GC} \in \{600, 900\}$ and both registered EH levels, training to 32,000 steps (128 arms total). The protocol and all analysis scripts were hash-bound before execution.

Prefix replay. All 128 refork arms reproduced their first 8,000 steps relative to the original fork-stage arms bitwise-exactly (max absolute difference 0; 128/128 `PREFIX_REPLAY_EXACT`), confirming state-matching fidelity.

Arrival outcome. All 128 extended-horizon arms reached H by 32,000 steps. The early low-GC arms therefore eventually arrive: the $8{,}000$-step completion separation reflects a difference in arrival *timing*, not a permanent gate.

Censor-aware AFT model. Combining the 128 refork records with the existing early ($q_{GC} = 1200$) and post-M three-dose arms yielded 288 records (287 observed, 1 censored). We fit a parent×stage fixed-intercept log-normal AFT model:

$$\begin{aligned}\log \tau_i = \alpha_{p(i),\,s(i)} + \beta_{E,\,s(i)}\, x_{E,i} \\ + \theta_{600}\,\mathbf{1}(q_{GC,i}{=}600) + \theta_{1200}\,\mathbf{1}(q_{GC,i}{=}1200) \\ + \delta_{600}\,\mathbf{1}(q_{GC,i}{=}600)\,\mathbf{1}(s(i) \in \text{early}) \\ + \delta_{1200}\,\mathbf{1}(q_{GC,i}{=}1200)\,\mathbf{1}(s(i) \in \text{early}) + \varepsilon_i,\end{aligned}$$

where τ_i is the time from the fork to high-accuracy confirmation, $x_{E,i} = (q_{EH,i} - 900)/300$, $\alpha_{p,s}$ is a parent×stage fixed intercept (implemented with a global intercept and one omitted parent×stage reference level), $\beta_{E,s}$ is a stage-specific EH nuisance slope whose columns multiply the stage indicators by x_E, q_{GC}=900 is the reference dose, and "early" denotes pre-M and M-onset stages pooled. The common-profile null ($\delta_{600} = \delta_{1200} = 0$) is rejected (LR= 53.03; inference below). The likelihood uses observed log-normal densities and right-censor survival probabilities at each arm's actual cap; early low/mid-GC caps are 32k, early high-GC and all post-M caps remain 8k, and exactly one high-GC observation remains censored.

Table 8: Post-hoc unified 8k completion comparison. Model entries are in-sample expected counts, not observed binomial counts. Early pools two stages and two EH levels for each of 16 parents. RD is GC1200 minus GC600 completion probability, in percentage points.

Stage	Data/model	GC600	GC1200	RD
Early	Observed	$0/64$	$63/64$	98.44
	M_0: common GC profile	$\approx 0/64$	$59.69/64$	93.27
	M_1: stage interactions	$\approx 0/64$	$61.66/64$	96.35
Post-M	Observed	$32/32$	$32/32$	0.00
	M_0	$30.71/32$	$32.00/32$	4.05
	M_1	$31.78/32$	$32.00/32$	0.69

Conditional null calibration. From the fitted null, we simulate independent log-normal times conditional on the fixed 16-parent design, apply the original arm-specific caps, and refit both models for each of 10,000 replicates. The observed LR$= 53.0282$ has zero exceedances, giving the finite Monte Carlo estimate $p_{\mathrm{MC}} = (0+1)/(10000+1)$; the one-sided 95% Monte Carlo upper bound for the underlying tail probability is 0.000300. This is a model-conditional calibration, not a cluster-robust population-level test.

Observed-parent uncertainty. For a separate 10,000-draw pairs bootstrap, we sample the 16 observed parents with replacement, retain all 18 outcomes and censoring indicators per sampled parent, relabel repeated copies, and refit the alternative model. This preserves observed within-parent dependence. These resamples do not impose the null, so their uncentered likelihood ratios are not used to calculate a null p-value. With 16 parents, percentile intervals are approximate and reported marginally, not simultaneously. All 20,000 replicates completed; no failed replicate was dropped.

Correction record. An earlier parent-block routine omitted the x_E multiplier in its EH nuisance columns and combined independent null simulation with parent resampling; its p-value is retired. The original observed AFT fit used the correct EH columns; the repaired shared implementation exactly reproduces its log likelihoods and LR before bootstrapping. The conditional calibration above and the observed-parent pairs intervals below replace the earlier block-bootstrap inference. This repair was specified after inspecting the existing outcomes and is labeled a corrective reanalysis in the artifact.

Interaction estimates. The point estimates are $\hat{\delta}_{600} = +0.1053$ and $\hat{\delta}_{1200} = -0.0629$, corresponding to approximately $+11\%$ and -6% median-time shifts at the two early GC levels. The observed-parent percentile 95% intervals are $[0.0646, 0.1559]$ and $[-0.0935, -0.0269]$, respectively. The preregistered materiality threshold is $|\delta| > \log(1.25) \approx 0.223$; both intervals lie within this band, so the interaction is detectable but not materially large.

Unified-horizon model description and sensitivity analysis. The common-GC-profile model M_0 retains parent×stage baselines and stage-specific EH slopes; M_1 adds the two early GC interactions. Table 8 evaluates both at a common 8k horizon while retaining actual arm-specific caps in fitting. Adding the interactions increases the early high-minus-low completion risk difference by 3.07 percentage points (parent-pairs 95% interval $[1.47, 4.66]$). The early-minus-post-M difference of risk differences increases from 89.23 to 95.66 points, an increment of 6.43 points ($[2.83, 10.74]$).

This post-hoc analysis uses the same 288 records (16 parents, 18 arms per parent; 287 events and one right-censored arm) and exactly reproduces the saved M_0/M_1 coefficients, scales, and log likelihoods. It does not refit with artificial 8k censoring. For each observed design row i, the fitted log-normal model gives

$$\widehat{p}_{m,i}(8{,}000) = \Phi\left(\frac{\log(8{,}000) - X_{m,i}\widehat{\beta}_m}{\widehat{\sigma}_m}\right).$$

Expected counts sum these probabilities over the specified stage/dose cells, averaging the two EH levels equally. Observed completion is $\mathbf{1}(T_i \le 8{,}000)$; all actual caps are at least 8k, so its status is known for every arm. The zeroed-interaction sensitivity retains the M_1 design and nuisance

coefficients and $\widehat{\sigma}_1$, setting only δ_{600} and δ_{1200} to zero. It is not a fitted null. Post-M and reference-dose predictions are therefore unchanged by this operation. The zeroed-interaction early expected counts are $0.0011/64$ at GC600 and $57.02/64$ at GC1200, giving an RD of 89.09 percentage points. At the unified 8k horizon, zeroing the two interactions reduces early RD by 7.25 percentage points (parent-pairs 95% interval $[2.57, 12.86]$), distinct from the 3.07-point difference between the refitted models.

For uncertainty, 10,000 parent-pairs bootstrap draws (seed 20260909) resample the 16 parents with replacement, retain all 18 original outcomes and censor indicators per copy, relabel duplicate parents, and refit both models. Predictions and paired differences are evaluated on each resampled cohort. All draws completed without discarded fits or optimizer retries. The reported percentile intervals are approximate marginal 95% intervals, not simultaneous or prediction intervals; they do not impose a null. The early RD intervals for M_0, M_1, and zeroed M_1 are respectively $[88.52, 97.48]$, $[92.97, 99.04]$, and $[81.85, 95.21]$ percentage points. For early-minus-post-M RD they are $[81.65, 95.75]$ for M_0 and $[91.79, 98.78]$ for M_1. The paired difference intervals reported above are calculated within each draw, not by subtracting marginal interval endpoints.

The high-versus-low GC cross-stage interaction is the distinct contrast $\delta_{1200} - \delta_{600} = -0.168$ (joint-draw interval $[-0.204, -0.132]$), corresponding to a time ratio of 0.845 ($[0.815, 0.876]$). It is not either individual reference-dose interaction or a new joint equivalence test. The registered log band $\pm \log(1.25)$ corresponds to multiplicative ratios $[0.8, 1.25]$, not symmetric percentage changes. The original marginal interaction intervals should not be interpreted as a formal simultaneous equivalence conclusion.

The existing maximum cell event-rate discrepancy of 0.024 evaluates M_1 at the actual arm-specific caps; it is not this unified-8k comparison. The new expectations are in-sample model descriptions conditional on the observed cohort and specification, not independent predictive validation or a unique causal decomposition. The common profile still has stage-specific EH slopes and parent×stage intercepts; no claim of completely stage-invariant treatment response follows. The analysis directory `iclr2027_v56_aft_completion_v1` retains the formal scripts, SHA256 input/code bindings, all bootstrap metrics and sampled parent IDs, and audit results. No new training or checkpoint measurement is involved.

Model adequacy and sensitivity. The maximum cell-level event-rate gap is 0.024. Alternative AFT specifications—Weibull ($\hat{\delta}_{600} = 0.119$, $\hat{\delta}_{1200} = -0.070$), log-logistic ($0.083$, -0.073), and stage-specific-scale—give the same interaction directions with modestly varying magnitudes. A common-linear log-time slope gives LR$= 42.54$ ($k_{\text{common}} = -0.343$); a stage-specific-scale model gives LR$= 60.53$.

Provenance. All refork trajectories, manifests, protocols, analysis scripts, and registry bindings are frozen and committed (commit `54de8f4d`). A post-training serialization bug (relative-path error) caused the runner manifest to be automatically flagged `TECHNICAL_INVALID`; this flag records a provenance serialization issue only—all 128 arm trainings and artifact writes completed before the error, and no arm was retrained or replaced. Endpoints and the analysis manifest were deterministically recovered under the frozen parser and endpoint rules; all 128 recovered reforks reproduced their historical first-8k behavioral arrays bitwise (`bitwise_equal=true`, `max_abs_diff=0`).

M S_5 Cross-Task Transport Protocol and Results

Task. The task is permutation composition over S_5: inputs are ordered pairs of permutations (σ, τ) over $\{1, \ldots, 5\}$, with target $\sigma \circ \tau$. The label space has 120 classes. The data split procedure and proportions are identical to the $\mathbb{Z}_{97}$ experiment (30% train, 20% calibration; Appendix B); the architecture, optimizer, and parameter grouping are also identical.

Design. Twelve fresh S_5 parents (seeds 430001–430012) are trained from initialization under uniform weight decay to an eligible step-800 post-memorization checkpoint. In all 12 parents, the baseline memorization onset T_M (three-checkpoint window with train acc ≥ 0.99, test acc < 0.10, relative CE drift ≤ 0.50) occurs at step 400 ($10/12$) or 450 ($2/12$), placing the fork 350–400 steps

after T_M. The same four-cell pressure-matched grid is used: $(q_{\mathrm{GC}}, q_{\mathrm{EH}}) \in \{600, 1200\}^2$, yielding $12 \times 4 = 48$ arms. The fork procedure and state-matching controls follow the $\mathbb{Z}_{97}$ factorial. The implemented S_5 endpoint is the confirmation-gated onset score Y_C^g (Appendix C). The continuation horizon is 32,200 steps, with registered analysis cutoffs at absolute calendar step C_H=12,000 (primary) and C_H=32,000 (de-censoring), corresponding to elapsed continuation budgets $L_H = C_H - 800 = 11{,}200$ and 31,200 steps respectively.

Source-swap binding. The registered estimand is the equal-total diagonal swap $D{=}Y_H^{\mathrm{cal}}(600, 1200){-}Y_H^{\mathrm{cal}}(1200, 600)$, where positive D means the GC-heavy allocation arrives earlier (supporting GC dominance). The source contrast vector was derived from the existing $\mathbb{Z}_{97}$ 12-parent pressure-matched factorial (seeds 311, 312, 313, 314, 316, 317, 318, 320, 321, 322, 324, 326) using the same Y_H^{cal} coordinate at C_H=12,000. The derived vector is $(2350, 2300, 1850, 2150, 3400, 1250, 1400, 1450, 3350, 2050, 1300, 2350)$, all $12/12$ positive (mean 2100, exact one-sided p=0.0002). This binding was frozen (SHA `99c3914b`) and the source gate passed before any S_5 target arm was generated.

State-matching audit. All 48 arms passed the formal audit (`FORMAL_AUDIT_PASSED`, 0 failures). The signed-swap mismatch (relative deviation from exact q-matching) ranges over $[5.7 \times 10^{-4}, 6.4 \times 10^{-4}]$, well below the 10^{-3} threshold. Fork-local λ_{GC} ranges over $[0.51, 1.10]$ and λ_{EH} over $[0.19, 0.39]$.

Table 9: S_5 per-parent equal-total swap contrasts $D{=}Y(600, 1200){-}Y(1200, 600)$ (steps). Positive D supports GC dominance. Subscripts denote calendar cutoffs.

Parent	D_{12k}	D_{32k}
430001	+900	+5400
430002	+2350	+2350
430003	+800	+800
430004	+1250	+1250
430005	−1400	−1650
430006	+1250	+2350
430007	+1050	+1600
430008	+1900	+1900
430009	+500	+500
430010	+1400	+1400
430011	0	+1650
430012	+2750	+2700
Mean	+1062.5	+1687.5
Pos./Neg./Zero	10/1/1	11/1/0
p (exact mean sign-flip)	0.005	0.002
p (concordance)	0.006	0.003
Bootstrap 95% CI	[450, 1621]	[813, 2592]

Endpoint rule and post-hoc sensitivity. The tables retain the original confirmation-gated onset scores, $Y_H^{\mathrm{cal}} = Y_C^g$ from Equation 7. Parent 430012 illustrates the boundary: the EH-heavy onset is 11950 but confirmation is 12150, while GC-heavy onset/confirmation is 9250/9450. The 12k score is therefore $12000 - 9250 = 2750$; the 32k score is $11950 - 9250 = 2700$. This decrease follows from the implemented gate and is not a property of ordinary truncation of a single event timestamp.

Using confirmation timestamps for both arms, Y_C^c, gives a post-hoc 12k mean contrast of 1012.5 steps (exact one-sided parent sign-flip $p = 0.00390625$; parent-bootstrap 95% interval $[441.7, 1545.8]$; 10 positive, one negative, one zero). At 32k, every pair is observed and adding 200 steps to both arms cancels: the contrast remains 1687.5 steps with unchanged paired inference. The 12k original analysis remains primary; this sensitivity documents the endpoint dependence, without changing the source binding or selecting a new primary cutoff. See the reproducible audit in Appendix C.

Censoring structure at $12k$. At C_H=12,000, 6 parents have both swap arms observed, 5 have one censored, and 1 (seed 430011) has both censored. At C_H=32,000, all 12 parents are fully observed.

Table 10: S_5 descriptive 2×2 cell means of Y_H^{cal} (calendar steps, fork at step 800). Lower = earlier arrival. Subscripts denote calendar cutoffs.

Cell $(q_{\text{GC}}, q_{\text{EH}})$	Y_H^{cal} at $C_H{=}12k$	Y_H^{cal} at $C_H{=}32k$
$(600, 600)$ (LL)	11733	14538
$(600, 1200)$ (EH-heavy)	10612	11300
$(1200, 600)$ (GC-heavy)	9550	9612
$(1200, 1200)$ (HH)	8229	8229

Interpretation. The cell ordering HH $<$ GC-heavy $<$ EH-heavy $<$ LL at both horizons is descriptively consistent with the $\mathbb{Z}_{97}$ finding that higher total pressure accelerates arrival. However, only the equal-total diagonal swap (GC-heavy vs. EH-heavy) is a registered inferential test; the full 2×2 table is descriptive and is not used for formal inference.

Horizon alignment. Because the S_5 fork occurs at step 800, the primary absolute cutoff $C_H{=}12{,}000$ provides 11,200 elapsed continuation steps—3,200 more than the $\mathbb{Z}_{97}$ stage experiment's $L_H{=}8{,}000$. Table 11 reports cumulative incidence at $L_H{=}4{,}000$, 6,000, and 8,000 elapsed continuation steps, matching the $\mathbb{Z}_{97}$ stage-experiment horizons. At the matched $L_H{=}8{,}000$, S_5 arrivals remain far below the $\mathbb{Z}_{97}$ post-M ceiling ($32/32$ per cell): the S_5 grokking time scale is approximately $2.1\times$ slower (per-cell median ratio 1.76–2.34, coefficient of variation 0.10). Descriptive calendar-time cell means (Table 10) retain the ordering HH $<$ GC-heavy $<$ EH-heavy $<$ LL in both tasks; however, at the matched $L_H{=}8{,}000$ the $\mathbb{Z}_{97}$ post-M cells have all reached the endpoint while S_5 remains substantially censored, so Table 11 is a descriptive time-scale comparison, not an effect-magnitude transport test. The formal transport inference uses the common calendar cutoff $C_H{=}12{,}000$ ($L_H{=}11{,}200$) for both tasks and does not rely on Table 11.

Table 11: S_5 cumulative incidence at $\mathbb{Z}_{97}$-matched elapsed continuation horizons. The S_5 fork is post-M ($\Delta{=}350$–400); the $\mathbb{Z}_{97}$ post-M fork has $\Delta{=}450$.

	S_5 (post-M)			$\mathbb{Z}_{97}$ (post-M)
Cell	$L_H{=}4k$	$L_H{=}6k$	$L_H{=}8k$	$L_H{=}8k$
LL ($q_{\text{GC}}{=}600$)	0/12	0/12	0/12	16/16
C2 ($q_{\text{GC}}{=}600$, $q_{\text{EH}}{=}1200$)	0/12	0/12	1/12	16/16
C1 ($q_{\text{GC}}{=}1200$)	0/12	1/12	4/12	16/16
HH ($q_{\text{GC}}{=}1200$, $q_{\text{EH}}{=}1200$)	0/12	3/12	8/12	16/16

Post-fork norm trajectory audit. A post-hoc within-parent audit of log-norm paths $R_g(t){=}\log\|\theta_g(t)\|/\|\theta_g(800)\|$ confirms reproducibly distinct coarse norm paths: in all 12 parents, GC-heavy contracts GC more and EH less than EH-heavy ($12/12$ sign-concordant). The separation deviates from pure-decay prediction $R_g^{\text{decay}}{=}L\log(1{-}\eta\lambda_g)$: GC retains 2–25%, EH 23–90%. This asymmetric attenuation reveals a non-isomorphism—the GC arrival advantage is not explained by larger realized GC displacement; at T_H the EH/GC ratio spans 1.08–2.50 with no unique norm endpoint. These paths are descriptive, not mediational.

N History-Conditioned Parameter-State Transplant: Protocol and Provenance

Design. The transplant factorial prospectively crosses two factors in 32 fresh $\mathbb{Z}_{97}$ parents (seeds 133001–133032). Each parent contributes four arms: $A{=}Y[\theta_{\text{G}}, O_{\text{G}}]$, $B{=}Y[\theta_{\text{G}}, O_{\text{E}}]$, $C{=}Y[\theta_{\text{E}}, O_{\text{G}}]$, $D{=}Y[\theta_{\text{E}}, O_{\text{E}}]$, where θ is the parameter configuration source (GC-dominated vs. EH-dominated early decay history) and O is the optimizer-history source (G-history vs. E-history Adam state). All four cells are prospective; the source pool was locked before any outcome was observed. The same common continuation policy, RNG state, and batch-stream continuation are shared across cells within a parent. The two source factors are assembled independently: model parameters are taken from the designated θ-source, while the complete Adam historical state (moments and associated carried optimizer state) is taken from the designated O-source.

Primary effect. Transplant timing uses episode onset with confirmation look-ahead (rule o, Appendix C); all 128 arms are observed, so the contrasts are unchanged if each event is timestamped 200 steps later. The parameter-source main effect is $\bar{\tau}_\theta{=}-810.16$ steps ($31/32$ strict concordance; two-sided exact sign-flip $p{=}9.31{\times}10^{-10}$). Both simple effects are concordant: $A{-}C{=}-895.31$ ($32/32$) and $B{-}D{=}-725.00$ ($31/32$). All preregistered gates passed ($n{=}32$, 500-step materiality, $24/32$ concordance, $p{<}0.05$).

Optimizer and interaction effects. The optimizer-source effect $\bar{\tau}_O{=}-41.41$ is not significant ($p{=}0.365$). The interaction $I{=}-170.31$ (Holm $p{=}0.042$) fails both the 500-step materiality gate ($|I|{<}500$) and the $24/32$ concordance gate ($21/32$), so material context dependence is not established.

Compatibility calibration. The native-only calibration compared each native cell with same-cell records from other parents. Across 64 native records, candidate thresholds $\kappa = 2, 4, 6, 8$ produced 27, 20, 14, and 10 false crossings, respectively; none satisfied the frozen zero-false-crossing criterion. Consequently no κ was selected and hybrid compatibility adjudication was not run. This is failure to obtain a compatibility certificate, not positive evidence that hybrid states are incompatible. The four-cell operational estimand remains reported, but component-carrying, mediation, stable parameter-source mechanism, and optimizer-robustness claims are unsupported. Source: `confirmation_kappa_calibration_v2.2.json`. No equivalence test for the optimizer-source main effect was registered; its nonsignificance does not establish practical equivalence to zero.

Provenance. The 32 parents, 128 arms, 5 formal artifacts, and protocol are frozen. Protocol SHA-256: `bad7f478`. An earlier non-formal attempt ($50/50/100$ steps vs. registered $450/450/8200$) was detected and aborted before any formal results were produced; formal results from the aborted run are not reused. Source data: `factorial_analysis_v2.2.json`.

O WIDTH-ROBUSTNESS: $d_{\text{model}}{=}96$ COMMON-CORNER REPLICATION

A formal replication at $d_{\text{model}}{=}96$ ($d_{\text{ff}}{=}384$, 4 heads, 2 layers) uses 18 fresh parents in a common four-corner design with coefficient-matched GC and EH decay levels ($\{g_{\text{low}}, g_{\text{high}}\} \times \{e_{\text{low}}, e_{\text{high}}\}$, 72 arms total, fork step 800, arrival cutoff 12,000, endpoint 18,000). This is a coefficient-local replication, not a pressure-matched experiment: the corner levels are set by raw decay coefficients rather than the registered q_g coordinate.

The GC arrival contrast δ_{GC} (high-minus-low, with EH averaged) is positive in $18/18$ parents (exact sign-flip $p{=}3.8{\times}10^{-6}$). The mean GC contrast is 2,378 steps; the mean EH contrast is 547 steps (ratio 4.35), confirming GC dominance in direction and magnitude. The GC-minus-EH interaction contrast is also positive in $18/18$ parents (Holm-adjusted $p{=}7.6{\times}10^{-6}$). For comparison, the same common-corner design at $d_{\text{model}}{=}64$ gives ratio 1.87 ($\delta_{\text{GC}}{=}2{,}248$, $\delta_{\text{EH}}{=}1{,}205$; $I_D{=}0.067$), and the original $d_{\text{model}}{=}128$ corners give ratio 5.80 ($\delta_{\text{GC}}{=}2{,}538$, $\delta_{\text{EH}}{=}438$; $I_D{=}0.833$). The GC-dominated direction is robust across widths; the magnitude ratio is width-dependent, consistent with the coordinate-sensitivity reported in Appendix G.

P TASK-ROBUSTNESS: MULTIPLICATION COEFFICIENT-LOCAL REPLICATION

A $3{\times}3$ coefficient-matched factorial on $\mathbb{Z}_{97}$ *multiplication* (target $\sigma \cdot \tau \bmod 97$, 97-class) uses 8 parents (72 arms, 0 censored). The GC high-minus-low arrival contrast is 2,000 steps, compared with 1,931 steps for the corresponding addition factorial. The EH high-minus-low contrasts are -131 steps (multiplication) and -100 steps (addition), both within the preregistered ± 300-step equivalence margin. This is a coefficient-local replication on a different algebraic operation within the same modulus, not a pressure-matched design. It confirms that the GC arrival contrast ordering replicates under a structurally related task, consistent with—but not identical to—the S_5 cross-task transport result (Section 6).

Q PROSPECTIVELY CONFIRMED LOCAL RESPONSE GEOMETRY: PROTOCOL AND RESULTS

Cohort and registration. Seeds 134001–134028 were preregistered for a confirmatory (not discovery) experiment and verified disjoint from every prior cohort by a filename/text exposure scan of the local and backup experiment trees. The protocol passed three external review rounds and had its design frozen before fresh training; a separate launch authorization bound the design and audit-bundle hashes. State, write, and measurement freezes were then created sequentially after their respective construction stages, with the measurement freeze preceding the single-table decision. Of 28 trained candidates, the first 24 eligible in ascending seed order formed the family (no replacement; all 28 records retained). All 96 cells reached the registered write-tangent qualification under frozen numerical gates (paired FP32/FP64 tangent concordance $\leq 10^{-4}$; finite-difference locality validation per state).

Assay. The assay uses cell labels A, B, C, D for the source combinations GG, GE, EG, EE, respectively (the combinations indexed explicitly in Section 7); the cell label B is distinct from the write-vector symbol. At each transplant cell, the registered source-normalized pulse defines the write vector B; the output tangent $K = J_Z B$ is evaluated on the full $9{,}409$-input centered-logit readout by forward-mode differentiation, with paired-precision replay and bitwise reproduction gates. Cells sharing a parameter source share B (bitwise), giving the exact decomposition $T^\theta = D_J + D_B$ of the main text with $c_J + c_B = 1$ (verified to 4.4×10^{-16} on the burned discovery cohort; the invariant is exact by construction).

Frozen decision rule and fresh results. The rule was frozen before training: H1 requires median $c_J^{\mathrm{swap}} \geq 0.75$ and $\geq 22/24$ parents with $c_J > 0.5$; H2 requires median $b_p \leq 0.70$, median $k_p \geq 0.95$, and $\geq 22/24$ parents jointly satisfying $b_p \leq 0.70 \wedge k_p \geq 0.95$ (all quantities FP64; family median $(x_{(12)}+x_{(13)})/2$). With signed contributions $c_J = \langle D_J, T^\theta\rangle / \|T^\theta\|_2^2$ and c_B defined analogously, $c_J + c_B = 1$ identically (values above 1 indicate partial cancellation, not explained variance). The registered per-parent statistics are $b_p = \max_{s\in\{\mathrm{G,E}\}} |\cos(B_s^{\mathrm{swap}}, B_s^{\mathrm{total}})|$ (non-collinearity in parameter space, absolute value so that anti-parallel write vectors do not count as distinct axes) and $k_p = \min_{X\in\{A,B,C,D\}} \cos(K_X^{\mathrm{swap}}, K_X^{\mathrm{total}})$ (worst-cell functional parallelism). The single-table decision (executed once, reading only the hash-bound table) returned: median $c_J = 0.9375$ (min 0.8287, max 1.0948), $24/24$ with $c_J > 0.5$; median $b_p = 0.6152$ (max 0.6392), $24/24 \leq 0.70$; median $k_p = 0.9947$ (min 0.9856), $24/24 \geq 0.95$; joint concordance $24/24$. Status: CONFIRMED. Signed contributions above 1 (one parent, $c_J = 1.095$) indicate partial cancellation by the write-vector term, and are reported as signed projections, not explained variance.

Boundary. The result confirms a one-step algebraic decomposition on the registered swap axis within the registered eligibility stratum; it does not establish mediation of the arrival effect, behavioral prediction, persistence beyond one step, or any state-dimension claim, and it shares no cohort with the transplant experiment.

R FROM LOCAL RESPONSE TO ARRIVAL: AN OPEN MECHANISTIC QUESTION

Differential persistence of decay-induced perturbations is one possible link between local response geometry and arrival timing. Its direction and magnitude may depend on the joint parameter–optimizer state and the subsequent training trajectory. The one-step assay in Section 7.2 characterizes the initial response; establishing a mechanism requires connecting its evolution under continued optimization to later arrival, rather than assuming a uniform GC-over-EH persistence advantage.

The existing mechanism analyses delimit this question: the specific $(k{=}1)$ Fourier mediator is unsupported, global GC-norm shrinkage is unnecessary in the observed embedding-only examples, and a replicated harmonic-marker response (GC > EH, $24/24$, $p \approx 6 \times 10^{-8}$) does not establish mediation (Appendix T). A direct behavioral bridge would require relating the evolution of controlled perturbations to subsequent arrival while accounting for the inherited training state. The transplant and one-step geometry results provide complementary observations for that question.

S PREDICTIVE-STATE BOUNDARY AUDIT

A frozen retrospective audit of the 32-parent transplant checkpoints (Section 7) examined whether registered checkpoint summaries—parameter-geometry, Adam-moment, representation, gradient, and curvature features—predict parent-specific parameter-source-effect heterogeneity $\tilde{\tau}_{\theta,p}$. Source identity is decodable out of parent from registered checkpoint features (factor-label LOPO R^2=0.295), showing that G- versus E-history leaves measurable checkpoint imprints. However, no registered scalar, supervised one-index, low-dimensional nonlinear readout, or full 32-feature dictionary achieved held-out prediction of $\tilde{\tau}_{\theta,p}$ heterogeneity (Table 12). This is a registered assay failure, not a proof of nonexistence: the result does not exclude an unmeasured scalar state, and the $\tilde{\tau}_{\theta,p}$ values themselves (mean -852, SD 388) contain continuation noise that limits detection power at n=32.

Table 12: Predictive-state audit results. Panel A: source identity is decodable from checkpoint features but is a distinct target from effect-heterogeneity prediction. Panel B: for the distinct target of parent-specific parameter-source-effect heterogeneity $\tilde{\tau}_{\theta,p}$, all registered predictors have negative held-out LOPO R^2; Axis-A reports incremental ΔR^2 beyond the source-label reference.

Panel A: Source-identity decoding (reference)	
Source-label baseline	0.295
Panel B: Effect-heterogeneity prediction	
Best Axis-A scalar, incremental ΔR^2 beyond source-label reference	-0.015 (95% CI $[-0.045, 0.136]$)
Best observed scalar	-0.008
Ridge 1-D learned index	-0.096
KRR on first 3 PCs	-0.494
Full 32-feature dictionary	-0.302
Raw $\Delta\theta$ RBF-KRR benchmark	-0.245

T MECHANISM EVIDENCE BOUNDARY

This appendix combines frozen post-hoc exclusion audits with an independent preregistered prospective confirmation of the treatment-to-marker response. None directly intervenes on a candidate mediator, so none establishes a mechanistic mediation claim.

Audit 1: Specific Fourier-$(k$=1$)$ mediator hypothesis is not supported. A descriptive Fourier trajectory analysis probed whether GC decay accelerates arrival by promoting alignment of the penultimate representation with the k=1 additive character of $\mathbb{Z}_{97}$ (Nanda et al., 2023). Terminal alignment is bimodal across 8 parents: $2/8$ reach $R^2{\approx}0.88$, while $6/8$ remain below 0.04—yet all 8 generalize to perfect test accuracy. The parent-blocked sign test for GC's effect on pre-arrival Fourier R^2 (step 3000) is $5/3$ (p=0.727); the per-arm difference is driven entirely by the two high-alignment parents. Within those two parents, GC is strongly associated with early R^2 (descriptive r=0.85), but R^2 has little residual association with onset after conditioning on GC (partial r=0.19). Because n_{parent}=2, this is descriptive only: we found no evidence that this specific Fourier score mediates the GC–arrival relationship, but cannot exclude mediation by richer Fourier structure. The result rules out the simplest version of "GC promotes grokking via k=1 Fourier basis formation," not all Fourier-based accounts.

Audit 2: Global GC-norm shrinkage is not necessary for arrival. One might hypothesize that EH decay helps less because it fails to shrink GC parameters, which are the ones that matter. A descriptive comparison of norm dynamics across routes contradicts the global version: in embedding-only arms (emb_wd=2.0, no GC decay), GC-norm *grows* by $+96\%$ to $+169\%$ ($\bar{\Delta}$=$+130\%$, n=5) between fork and arrival, while EH-norm contracts by -40% to -50%. In GC-decay arms, GC-norm contracts by -42% to -53% ($\bar{\Delta}$=-47%, n=12). Thus generalization can arrive with GC-norm growth, ruling out "global GC-norm shrinkage is necessary for arrival." This does not exclude mechanistic relevance of subgroup-specific or relative norm dynamics. These cohorts differ in seed pool and horizon (32k vs. 18k), so this is descriptive, not a paired comparison.

Audit 3: Differential harmonic-marker response replicates, but marker-to-arrival mediation is not established. Two independent cohorts confirm that GC and EH decay produce measurably different harmonic-marker responses at a pre-arrival landmark (step 2400). An initial 10-parent cohort found harmonic anisotropy D_H=+0.00516 (10/10 positive, one-sided exact p=0.002). A 24-parent confirmation replicated this: D_H=+0.0121 (24/24 positive, one-sided p=5.96×10^{-8}). This establishes:

$$q_{\mathrm{GC/EH}} \longrightarrow \text{different harmonic-marker response} \quad \checkmark$$

However, the incremental prediction from harmonic score to remaining arrival time was *not* established: both Family A (remaining-time prediction, p=0.42) and Family B (remaining-time prediction, p=0.16) failed their gates. The mediation link therefore remains open:

$$\text{harmonic marker} \longrightarrow T_H \quad \text{(not established)}$$

Summary of evidence boundary.

Candidate mechanism	Status	Evidence
Specific (k=1) Fourier-score mediator	Not supported	5 pos./3 neg., p=0.73; 6/8 grok with R^2<0.04
Global GC-norm shrinkage (necessary route)	Descriptively contradicted	GC-norm +130% in emb-only arms
Differential harmonic-marker response	Supported (replicated)	24/24, p≈6×10^{-8}
Harmonic marker → arrival mediation	Not established	Family A/B prediction gates failed
Core vs. embedding effect gap	Unresolved	Δ_{core}=1727, Δ_{emb}=969, p_{gap}=0.194

U WHOLE-TRAINING POLICY TRANSPORT

A prospectively fixed, unselected 16-seed initialization cohort (seeds 201–216) tests whether the GC-over-EH ordering persists when group-selective AdamW decay is applied from initialization rather than at a fork checkpoint. For each seed, four runs are trained from initialization: GC-only decay at $w_d \in \{0.5, 2.0\}$ and EH-only decay at $w_d \in \{0.5, 2.0\}$, yielding $16 \times 4 = 64$ arms with a policy horizon of 18,000 steps. Unlike the fork experiments, this design uses matched raw AdamW coefficients—not pressure-matched coordinates—and does not share a common parent checkpoint. All 64 arms completed (0 technical invalids); raw trajectories were re-derived and hash-bound to the formal-run manifest.

The registered arrival score is $Y_H = (1 - T_H/18000)_+$, where T_H is the first step achieving test accuracy ≥ 0.90 for 5 consecutive checkpoints. For dose d, let $D^H_{p,d} = Y^H_{p,\mathrm{GC},d} - Y^H_{p,\mathrm{EH},d}$; the registered seed-level policy contrast is $D^H_p = \frac{1}{2}(D^H_{p,0.5} + D^H_{p,2.0})$. The primary result is $\bar{D}_H$=0.141 (16/16 seed-level contrasts positive, exact sign-flip p=1.53×10^{-5}), supporting dose-averaged transport of the GC-over-EH arrival ordering to whole-training raw-coefficient policies. The hierarchical stable-onset score (requiring 10 consecutive checkpoints above threshold) also favors GC on average ($\bar{D}_S$=0.082, p=6.1×10^{-5}), but this result is dose-dependent.

The dose decomposition reveals a sharp asymmetry. At w_d=0.5, GC-only arms reach arrival in 15/16 seeds while EH-only arms reach 0/16—a finite-budget accessibility separation qualitatively similar to the early-fork result ($\bar{D}^H_{0.5}$=0.231). At w_d=2.0, the registered arrival score remains positive on average ($\bar{D}^H_{2.0}$=0.051), but this reflects earlier rare GC successes rather than higher completion incidence: GC-only arms arrive in 3/16 seeds (T_H=3,350, 3,900, 17,200) while EH-only arms arrive in 4/16 (all $T_H \in$ [11,700–16,700]). The seed-level high-dose contrast is 3 positive, 10 ties, 3 negative. Stable-onset ordering at w_d=2.0 is descriptively reversed ($\bar{D}^S_{2.0}$=−0.031; GC 1/16 vs EH 3/16). No independent high-dose confirmatory hypothesis was registered; the reversal is reported descriptively.

This raw-coefficient experiment is not pressure-matched and lacks uniform/no-decay benchmarks, so it establishes no practical advantage over standard AdamW. The cohort was prospectively fixed within the audited workspace; project-wide freshness is not claimed. GC contains ≈ 94% of parameters, but contraction depends on $\lambda_g\|\theta_g\|^2$, not count alone. Equal coefficients need not match initial or cumulative contraction.