BEYOND EMBEDDING TRANSFER: COMPONENT ROLES IN GROKKING TRANSFER AND STABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Warm-start transfer can make algorithmic tasks generalize rapidly, yet it is unclear which model components provide the gain and whether that gain remains stable under continued optimization. We study cross-operator transfer on modular arithmetic and separate *efficacy* (early velocity) from *stability* (post-reach drawdown). In a scale-matched 108-run battery across 12 seed blocks (a 96-run 2^3 factorial plus a 12-run scale control), transferring internal attention/MLP weight matrices (B) in addition to token embeddings and readout ($E + U$) improves early mean accuracy by 5.46 percentage points (Holm $p = 0.0039$) and reduces confirmation latency by 558 steps (Holm $p = 0.0088$). While readout plus internal-block transfer satisfies the pre-specified ± 500-step mean-latency equivalence criterion in one-layer models (TOST $p = 0.0011$, although Full is modestly faster in 11/12 paired seeds), a prospective two-layer replication confirms the internal-block acquisition advantage (12/12 seeds, $+704.67$ integral units, $p = 4.88 \times 10^{-4}$) while revealing an architecture-dependent boundary: omitting donor embeddings falls 4475.6 integral units below Full transfer, far outside the pre-specified ± 250-unit equivalence margin. Continued target training, however, frequently triggers severe post-grokking relapse. Freezing transferred representation carriers (E, U) nearly eliminates offline relapse ($19.40\% \rightarrow 0.07\%$, Holm $p = 0.005859$). Online validation-triggered gating slashes True Max Drawdown from 22.06% to 0.60% on $2a + b$ ($p = 0.000488$), with prospective confirmations extending protection across affine, nonlinear quadratic, and two-layer targets (10.94–23.47 pp reductions), while distinguishing continual stabilization from static early stopping. In non-abelian S_5, unshielded transfer surges transiently (95.4% peak), but a prospective shielding cohort yields no confirmed benefit ($+0.15 \pm 1.14$ pp). These results establish a component-level dissociation between transfer acceleration and trajectory stability, and expose the empirical boundaries of parameter shielding.

1 INTRODUCTION

When trained on algorithmic and algebraic reasoning tasks with weight decay, overparameterized neural networks often exhibit *grokking*: training loss drops to zero almost instantly via memorization, yet held-out generalization is delayed by thousands or tens of thousands of optimization steps (Power et al., 2022; Liu et al., 2022; 2023). Mechanistic interpretability has established that for modular arithmetic, delayed generalization corresponds to the gradual formation of structured Fourier multiplication circuits and circular representation manifolds in embedding and attention weights (Nanda et al., 2023; Gromov, 2023; Varma et al., 2023).

Recently, Xu et al. (2025) demonstrated that initializing target models with input embeddings learned by a weaker model accelerates generalization on algorithmic tasks. In cross-operator transfer, however, focusing solely on input representations leaves open whether internal attention/MLP weight matrices (B) convey independent computational utility, and how warm-started models evolve during continued optimization. In practice, unshielded fine-tuning frequently exhibits acute post-reach generalization relapse, where initial gains are disrupted rather than monotonically refined. This leads to three concrete questions:

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.

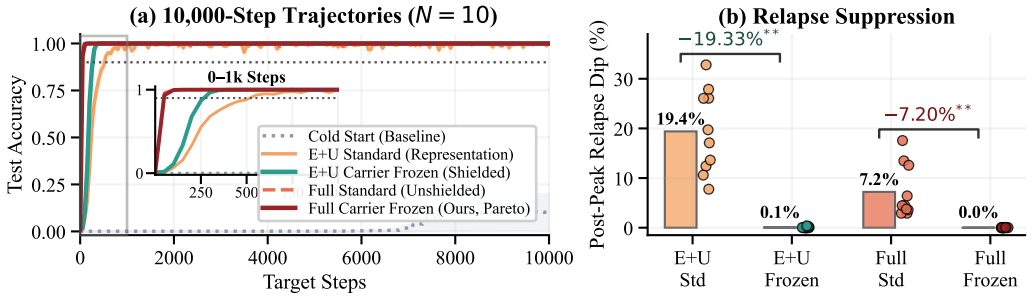

Figure 1: **Separating takeoff velocity from trajectory stability** ($N = 10$ **Blocks 230–239: 60-run confirmatory matrix plus 10-run exploratory extension).** (a) Test-accuracy trajectories ($\pm$ SEM). Full transfer reaches 90% by step 50, versus 250–500 steps for representation conditions. (b) Freezing $E + U$ reduces post-peak relapse from 19.40% to 0.07% (Holm $p = 0.005859$). The exploratory `full_carrier_frozen` extension combines 50-step takeoff with 0.02% relapse.

1. **Acquisition**: Do internal attention/MLP weight matrices (B) provide computational acceleration beyond representation carriers (E, U), and does this requirement change across architecture depths?
2. **Relapse**: Why do warm-started models relapse, and which components undergo contemporaneous degradation during post-reach generalization collapse?
3. **Protection**: Can targeted parameter-level interventions suppress relapse without degrading early velocity or terminal accuracy, and where do these protections break down?

Across these questions, our central insight is that **transfer acceleration and trajectory stability are governed by distinct component roles**: internal attention/MLP weight matrices (B) contribute strongly to acquisition velocity, while shielding representation carriers controls retention stability; the carrier requirements for acquisition become architecture-dependent with depth. Our contributions are:

1. **Component Roles and Architecture Dependence**: Scale-matched factorial decomposition shows that internal-block transfer provides substantial incremental acquisition acceleration (+5.46 pp early accuracy, Holm $p = 0.0039$; +704.67 units in 2-layer replication, $p = 4.88 \times 10^{-4}$); however, the role of donor embeddings changes with depth, where omitting donor embeddings in two layers causes a massive -4475.6-unit acquisition deficit.
2. **Acquisition and Stability as Distinct Lifecycle Properties**: Fast warm-start acquisition does not guarantee retention under continued optimization. On a probed relapse, factorial state replacement localizes the contemporaneous functional deficit primarily to internal blocks, while linear readout refitting recovers 8.37–11.03 points at troughs, demonstrating substantial preserved linear decodability during apparent generalization collapse.
3. **Validation-Triggered Carrier Protection and Its Boundaries**: Representation carrier ($E + U$) shielding strongly suppresses offline relapse ($19.40\% \to 0.07\%$, Holm $p = 0.005859$) and online validation gating slashes max drawdown from 22.06% to 0.60% on $2a + b$ ($p = 0.000488$) with consistent gains on affine ($+23.47$ pp) and quadratic ($+10.94$ pp) tasks. Crucially, we identify its empirical boundaries: shielding achieves confirmed attenuation but incomplete protection in two-layer models ($+15.35$ pp reduction, residual drawdown 65.23%), and yields no confirmed benefit in non-abelian S_5 ($+0.15 \pm 1.14$ pp).

2 Problem Formulation & Dual-Track Protocol

2.1 Task Definitions & Architecture

We investigate algorithmic transfer across two canonical algebraic settings: (1) **Abelian Modular Arithmetic** ($p = 113$): The source donor task is modular addition $a + b \pmod{113}$; the target task is the affine modular target $2a + b \pmod{113}$. The sequence format is $[a, b]$, predicting label $c \in \{0, \ldots, p - 1\}$. The dataset comprises $113 \times 113 = 12{,}769$ pairs, partitioned into 20% train ($N_{\text{train}} = 2{,}554$), 10% validation ($N_{\text{val}} = 1{,}277$), and 70% test ($N_{\text{test}} = 8{,}938$) via deterministic rounding. (2) **Non-Abelian Symmetric Group** (S_5): Source task is group multiplication $a \circ b$; target task is conjugation $a \circ b \circ a^{-1}$. Sequences format two operand tokens $[a, b]$ where $a, b \in \{0, \ldots, 119\}$,

predicting permutation index $c \in \{0, \ldots, 119\}$. The domain contains $120 \times 120 = 14{,}400$ pairs, partitioned deterministically into 30% train ($N_{\text{train}} = 4{,}320$), 20% validation ($N_{\text{val}} = 2{,}880$), and 50% test ($N_{\text{test}} = 7{,}200$). Models are standard pre-LN Transformers (Power et al., 2022; Nanda et al., 2023) with $d_{\text{model}} = 128, n_{\text{head}} = 4, d_{\text{mlp}} = 512$. In full factor transfer (or full 2D-weight transfer, denoted `full` or $E + U + B$), the network transfers token embeddings (E), the readout matrix (U), and internal attention/MLP weight matrices (B) with per-tensor Frobenius norm matching. Positional embeddings, LayerNorm parameters, and MLP biases retain their native cold-initialized values (exact parameter tensor mappings and initialization conventions are detailed in Appendix B). Optimization uses AdamW (Loshchilov & Hutter, 2019) with base learning rate $\eta = 10^{-3}$, weight decay $\lambda = 1.0$, betas $\beta = (0.9, 0.98)$, and batch size 512. Computations in the confirmatory battery are executed in standard FP32 precision (with PyTorch AMP utilized during earlier exploratory sweeps). Under carrier freezing interventions, only token embeddings W_E and readout head W_U are frozen ($\nabla_\theta = 0$); positional embeddings, LayerNorms, and internal blocks remain fully trainable.

2.2 Per-Tensor Scale Matching (Decoupling Direction from Magnitude)

To ensure that transfer benefits reflect parameter *direction* rather than arbitrary initialization *scale*, all target initializations in seed block b are paired with a trained, qualified donor checkpoint W_b^{donor} and a native cold initialization W_b^{cold}. A source run qualifies as a donor checkpoint only upon achieving stable grokking, defined by our prospectively specified gate as maintaining validation accuracy $\geq 90\%$ across a continuous 1,000-step persistence window (21 consecutive evaluations at cadence $\Delta = 50$). The donor checkpoint evaluated at the exact completion of this 21-point persistence window is saved and used as the qualified donor, with donor optimization terminating immediately upon saving to prevent post-hoc peak selection. For any transferred parameter tensor θ, its direction is transferred while its Frobenius norm is rescaled to match the native cold initialization: $\theta^{\text{transfer}} = (\theta^{\text{donor}} / (\|\theta^{\text{donor}}\|_F + 10^{-12})) \cdot \|\theta^{\text{cold}}\|_F$. This controls for tensor-level Frobenius norm scaling differences, isolating parameter direction from overall tensor magnitude.

2.3 Dual-Track Evaluation Metric Protocol

Prior transfer studies typically emphasize whether transfer *accelerates* learning, while the stability of those gains under continued optimization is less often separated explicitly. We distinguish validation accuracy $A_{\text{val}}(t)$ from held-out test accuracy $A_{\text{test}}(t)$. Test accuracy is evaluated strictly for reporting and never controls triggering, optimization, hyperparameter selection, or confirmatory stopping decisions:

Efficacy Track (Early Velocity). We quantify efficacy by: (1) first reach step $t_{\text{reach}} = \min\{t : A_{\text{val}}(t) \geq 0.90\}$; (2) pre-5k accuracy integral $I_{5\text{k}} = \int_0^{5000} A_{\text{test}}(t)\, dt \in [0, 5000]$; and (3) full-trajectory mean test accuracy $\bar{A} = \frac{1}{T}\int_0^T A_{\text{test}}(t)\, dt$.

Stability Track (Retention & Relapse). We quantify stability by: (1) 21-point persistence confirmation ($H = 1{,}000$ steps at cadence $\Delta = 50$, requiring $A_{\text{val}}(t) \geq 0.90$ across all $K = 21$ consecutive evaluations). For each seed block b, the confirmation latency is T_b, truncated at the uniform experimental horizon $\tau = 30{,}000$ steps as $\min(T_b, \tau)$. Across a cohort of N blocks, the Restricted Mean Survival Time is estimated by the sample mean

$$\widehat{\text{RMST}}_\tau = \frac{1}{N}\sum_{b=1}^{N} \min(T_b, \tau), \tag{1}$$

assuming administrative censoring at the maximum budget τ; (2) post-peak relapse dip $\Delta_{\text{relapse}} = A_{\text{test}}(t_{\text{peak}}) - \min_{t > t_{\text{peak}}} A_{\text{test}}(t)$ (primary stability metric for offline Phase C; maximum drawdown is supplementary); and (3) for online streaming interventions (§5.4–5.5), models branch from an online validation trigger step:

$$t_{\text{trig}} = \min\{t : A_{\text{val}}(t - \Delta) \geq q \wedge A_{\text{val}}(t) \geq q\}, \tag{2}$$

requiring two consecutive validation evaluations above threshold q at cadence $\Delta = 50$ (with $q = 0.95$ for modular arithmetic; $q = 0.70$ for S_5). Over a pre-specified finite observation horizon H steps

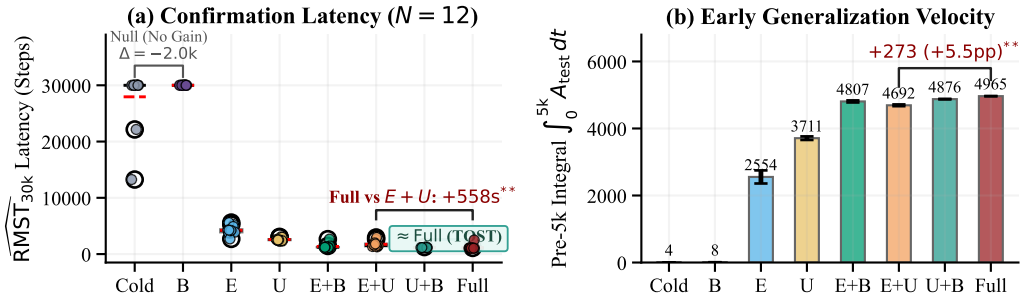

Figure 2: **Component Localization and Factorial Synergy ($N = 12$ Formal Blocks 210–221: 96 Factorial Runs plus 12 Scale Controls, 108 Runs Total).** (a) Confirmation latency $\widehat{\text{RMST}}_{30k}$ across 8 factorial conditions (medians and individual points). Full transfer accelerates confirmation by $+558.3$ steps over representation transfer $E + U$ (Holm $p = 0.0088$). Crucially, co-transferring readout and internal blocks ($U + B$) achieves TOST equivalence to full transfer ($p = 0.0011$, $\Delta = -16.7$ steps), while internal blocks alone (B) do not establish a confirmation benefit within the 30k horizon ($p = 0.50$). (b) Pre-5k accuracy integral, confirming that full transfer provides $+5.46$ pp higher early accuracy than $E + U$ (Holm $p = 0.0039$).

post-trigger ($H = 1{,}500$ for modular cohorts; $H = 3{,}000$ for S_5), we evaluate trajectory stability via the post-trigger True Max Drawdown:

$$D_H^{\text{trig}} = \max_{t_{\text{trig}} \le t \le t_{\text{trig}}+H} \left[\max_{t_{\text{trig}} \le s \le t} A_{\text{test}}(s) - A_{\text{test}}(t) \right]. \tag{3}$$

Because early-stopping halts at t_{trig}, its post-trigger drawdown is identically $D_H^{\text{trig}} \equiv 0$, matching Table 2.

2.4 Statistical Inference and Family-Wise Error Rate Control

All paired contrasts are evaluated across paired seed blocks using exact two-sided sign-flip permutation tests (2^N complete enumerations: $2^{12} = 4{,}096$ in Phase B; $2^{10} = 1{,}024$ in Phase C). In Phase B, family-wise error rate is controlled at $\alpha = 0.05$ via the step-down Holm-Bonferroni procedure (Holm, 1979) separately for the pre-specified eight-contrast efficacy family (I_{5k}, Table 1) and the pre-specified eight-contrast stability family (RMST_{30k}), spanning cold vs. individual components, component additions, and scale control, with confidence intervals estimated via 10,000-resample Bias-Corrected and Accelerated (BCa) bootstrapping. For the primary contrast (Full vs. $E + U$), exact two-sided sign-flip permutation tests ($2^{12} = 4{,}096$ enumerations) yield raw $p_{\text{perm}} = 2/4096 = 0.000488$ on efficacy; because the planned efficacy family contains $m = 8$ contrasts, the initial Holm multiplier is $m = 8$, yielding $p_{\text{Holm}} = 8 \times 0.000488 = 0.0039$. For confirmation latency, raw $p_{\text{perm}} = 0.001221$ yields step-down adjusted $p_{\text{Holm}} = 7 \times 0.001221 = 0.0088$. In the Phase C confirmatory battery, the superiority contrast family (carrier freezing, differential learning rate, temporary freezing) is evaluated under exact permutation tests with Holm-Bonferroni correction ($\alpha = 0.01$ pre-specified criterion) and 2,000-resample BCa bootstrapping, while non-inferiority (representation-only early velocity vs. full transfer) is evaluated via a one-sided shifted permutation test at $\alpha = 0.05$ against a pre-specified indifference margin ($\delta = 100.0$ integral units). Equivalence in Phase B is assessed using a paired Student-t Two One-Sided Tests (TOST) procedure (Schuirmann, 1987) against a pre-specified ± 500-step RMST margin (rather than a permutation test).

3 What Information Is Transferred? Carrier Localization and Synergy

To identify component-level contributions to transferred computation, Phase B executes a 108-run battery across 12 formal blocks (Blocks 210–221: a 96-run 2^3 factorial ablation plus a 12-run scale control) on $a + b \to 2a + b \pmod{113}$. We decompose the model into: token embeddings (E), readout classification head (U), and internal transformer blocks (B, including multi-head self-attention and MLP layers). Table 1 and Figure 2 report the complete findings.

Table 1: **Phase B Battery and Inferential Contrasts ($p = 113$, $N = 12$ paired blocks 210–221: 96 factorial runs plus 12 scale controls, 108 runs total). Panel A (Descriptive Factorial Summary)**: Confirmation fraction (ValAcc $\geq 90\%$ sustained across 1,000 steps), mean confirmation latency ($\widehat{\text{RMST}}_{30k}$), pre-5k accuracy integral (I_{5k}), early mean test accuracy, and descriptive paired difference vs. full transfer ($\Delta\text{RMST} = \text{RMST}_{\text{cond}} - \text{RMST}_{\text{full}}$; positive values indicate slower confirmation than Full). Uncertainties are Mean ± SEM ($N = 12$). The scale control condition (`embed_orig_scale`) achieves RMST $18{,}025.0 \pm 1{,}944.4$ steps, $I_{5k} = 1{,}131.6 \pm 208.5$. **Panel B (Prospectively Specified Inferential Contrasts and Equivalence)**: Formal statistical tests across paired blocks with explicit algebraic estimand formulas. Raw p_{perm} is from exact two-sided sign-flip permutation tests ($2^{12} = 4{,}096$ enumerations) with Holm-Bonferroni FWER control ($\alpha = 0.05$). For internal blocks alone (B vs. Cold), 10 of 12 blocks are mutually censored at $\tau = 30{,}000$ steps ($N_{\text{non-zero}} = 2$), so the non-crossing BCa CI reflects pervasive censoring rather than significant divergence from the exact null ($p = 0.5000$). Equivalence for $U + B$ vs. Full is evaluated via a paired Student-t Two One-Sided Tests (TOST) procedure (Schuirmann, 1987) against the pre-specified margin (±500 steps; $p_{\text{TOST}} = 0.001089$); the 12 paired differences comprise nine +100-step differences, one +50-step difference (Block 216), one +200-step difference (Block 218), and one −1,350-step difference (Block 217; mean −16.7 steps), establishing equivalence of average confirmation latency under the pre-specified tolerance rather than identical per-seed trajectories.

Panel A: Factorial Condition Summary ($N = 12$ paired blocks: 96 factorial runs plus 12 scale controls)

Condition	Confirmed	Mean RMST (Steps)	Mean I_{5k} (Units)	Mean Acc (0–5k)	Paired ΔRMST vs. Full [95% BCa CI]
Cold Start (`cold`)	2/12 (16.7%)	$27{,}950.0 \pm 1{,}486.7$	3.9 ± 0.2	0.08%	+26,775.0 [21392.8, 28700.0]
Internal Only (`B`)	0/12 (0.0%)	$30{,}000.0 \pm 0.0$	7.5 ± 0.2	0.15%	+28,825.0 [28321.1, 28945.8]
Embedding Only (`E`)	12/12 (100.0%)	$4{,}220.8 \pm 219.5$	$2{,}553.7 \pm 204.0$	51.07%	+3,045.8 [2541.7, 3520.8]
Readout Only (`U`)	12/12 (100.0%)	$2{,}600.0 \pm 66.0$	$3{,}710.6 \pm 55.7$	74.21%	+1,425.0 [1011.5, 1608.3]
Embed + Block (`E+B`)	12/12 (100.0%)	$1{,}354.2 \pm 119.7$	$4{,}806.9 \pm 32.6$	96.14%	+179.2 [−120.8, 516.7]
Representation (`E+U`)	12/12 (100.0%)	$1{,}733.3 \pm 153.7$	$4{,}691.8 \pm 27.4$	93.84%	+558.3 [270.8, 954.2]
Readout + Block (`U+B`)	12/12 (100.0%)	$\mathbf{1{,}158.3 \pm 8.3}$	$4{,}876.0 \pm 8.4$	97.52%	**−16.7** [−500.0, 108.3]
Full Factor (`full`, $E + U + B$)	**12/12 (100.0%)**	$\mathbf{1{,}175.0 \pm 120.5}$	$\mathbf{4{,}964.8 \pm 2.4}$	**99.30%**	Baseline (0.0)

Panel B: Prospectively Specified Inferential Contrasts and Equivalence Tests

Target Contrast	Contrast (A vs. B)	Algebraic Estimand Formula	Mean Difference [95% BCa CI]	Raw p_{perm}	Holm / TOST p
Internal Efficacy Gain	Full vs $E + U$	$I_{5k}(\text{full}) - I_{5k}(E + U)$	+273.0 [229.9, 339.7]	0.0005	$p_{\text{Holm}} = 0.0039$
Internal Latency Saving	Full vs $E + U$	$\text{RMST}_{E+U} - \text{RMST}_{\text{full}}$	+558.3 [270.8, 954.2]	0.0044	$p_{\text{Holm}} = 0.0088$
Overall Transfer Gain	Cold vs Full	$\text{RMST}_{\text{cold}} - \text{RMST}_{\text{full}}$	+26,775.0 [21392.8, 28700.0]	0.0005	$p_{\text{Holm}} = 0.0039$
Internal Blocks Alone	Cold vs B	$\text{RMST}_{\text{cold}} - \text{RMST}_{B}$	−2,050.0 [−8649.1, −654.2]	0.5000	$p_{\text{Holm}} = 0.5000$ (n.s.)
Latency Equivalence (TOST)	$U + B$ vs Full	$\text{RMST}_{U+B} - \text{RMST}_{\text{full}}$	−16.7 [−500.0, 108.3]	—	$p_{\text{TOST}} = 0.0011$

3.1 Internal Weight Matrices Provide Significant Incremental Acceleration

While transferring input and output representations (`E+U`) substantially accelerates confirmation relative to cold start (RMST 1,733.3 vs. 27,950.0), transferring the full factor direction ($E + U + B$) provides significant incremental advantages along both evaluation tracks: (1) **Efficacy Track**: Full transfer increases the pre-5k integral by +273.0 units (95% BCa CI: $[229.9, 339.7]$), representing an average test accuracy gain of $+\mathbf{5.46}$ **percentage points** over the first 5,000 steps (Holm $p = 0.0039 < 0.01$). (2) **Stability Track**: Full transfer reduces continuous confirmation latency by $+\mathbf{558.3}$ **steps** (95% BCa CI: $[270.8, 954.2]$, Holm $p = 0.0088 < 0.01$). Both tracks show a significant incremental benefit under FWER control: transferring internal attention/MLP weight matrices (B) improves learning beyond $E + U$ transfer alone.

3.2 Factorial Synergy: Readout Power and Internal Independence

Examining isolated component transfers reveals unexpected structural asymmetries:

- **The Readout Head Is a Strong Independent Carrier**: Transferring the readout head alone (`U`) yields an integral of 3,710.6 units and confirmation in 2,600.0 steps, dramatically outperforming transferring the input embedding alone (`E`: 2,553.7 units, 4,220.8 steps).
- **B Alone Fails to Improve Confirmation Within the 30k Horizon**: Transferring B alone does not improve confirmation over cold start within the 30k horizon (0/12 confirmed; RMST 30,000.0 vs. 27,950.0, exact sign-flip $p = 0.5000$). Crucially, 10 of 12 blocks experienced mutual administrative censoring at $\tau = 30{,}000$ steps, leaving only two non-zero paired differences ($-7{,}850$ and $-16{,}750$ steps). While the descriptive BCa CI ($[-8{,}649.1, -654.2]$) does not cross zero due to resampling predominantly from these two negative pairs, this does not overturn the non-significant pre-specified exact permutation test, reflecting pervasive censoring ($N_{\text{non-zero}} = 2$) rather than confirmed divergence.
- **Scale Control Validates Magnitude Normalization**: The scale control condition (`embed_orig_scale`) exhibits an RMST of $18{,}025.0 \pm 1{,}944.4$ steps (+13,804.2-step latency penalty vs. scale-matched `E`, Holm $p = 0.0039$; Table 1), demonstrating empirically that unnor-

malized transfer of learned embedding magnitudes severely hinders downstream optimization and validating our scale-matching protocol.

- **Non-Additive Interaction**: The descriptive 3-way interaction $I_{EUB} = F_{\text{full}} - F_{EU} - F_{EB} - F_{UB} + F_E + F_U + F_B - F_0$ is negative ($-3{,}141.9$ integral units, $p = 0.0005$). Because $I_{5\text{k}}$ is bounded above by 5,000, we interpret this as ceiling-compressed diminishing returns rather than a distinct mechanism.

3.3 Bounded Latency Equivalence of $U + B$ vs. Full

The most striking discovery from Phase B emerges from comparing `U+B` against `full`. Across the 12 paired blocks, individual differences $\text{RMST}_{U+B} - \text{RMST}_{\text{full}}$ comprise nine blocks with $+100$ steps (where Full confirms at 1,050 and $U + B$ at 1,150), one block with $+50$ steps (Block 216), one with $+200$ steps (Block 218), and a single block with $-1{,}350$ steps (Block 217, where Full experienced delayed confirmation at 2,500 while $U + B$ confirmed at 1,150). Across all 12 blocks, `U+B` achieves an average confirmation latency of $1{,}158.3$ steps (versus $1{,}175.0$ for full transfer), differing by $\Delta = -16.7$ steps (95% BCa CI: $[-500.0, 108.3]$). A paired Student-t TOST against the pre-specified margin $|\Delta| < 500$ steps rejects non-equivalence (using the conventional signs for the lower- and upper-bound tests, $t_L = 3.975$ and $t_U = -4.249$, $df = 11$, $p_{\text{TOST}} = 0.001089$), establishing operational equivalence of *average confirmation latency* within the pre-specified ± 500-step tolerance rather than computational parity or identical per-seed trajectories. The mean difference is influenced by one delayed Full-transfer block (Block 217). Excluding this block yields a mean $U + B -$ Full latency difference of $+104.5$ steps (median $+100$); the leave-one-out estimate remains comfortably within the pre-specified ± 500-step equivalence margin. Across the 12 seeds, 11 of 12 exhibit positive differences (two-sided sign test $p = 0.00635$). Thus, $U + B$ satisfies the pre-specified ± 500-step operational mean-latency equivalence criterion, although Full is modestly faster (by ≈ 100 steps) in 11/12 paired seeds; the average-equivalence conclusion is therefore not driven by Block 217. Within this setting and metric, early-integral equivalence was not established.

4 Why Does Generalization Relapse? Bounded Mechanistic Diagnostics

Continued target training can produce severe accuracy relapses after initially successful generalization. We combine diagnostics with distinct evidential scope: the state swap and displacement analyses examine an exploratory single probed trajectory (Blocks 210/230), while linear probes across four test-conditioned relapse events (three scratch, one transfer) characterize contemporaneous decodability deficits rather than establishing a universal causal mechanism. A single-block trajectory shows concurrent embedding-norm decline, readout-norm growth, and subspace drift; these do not track accuracy monotonically and are thus observations rather than causes (Appendix A.1). A 2^4 state-swap on one 25.24-point relapse exchanges embeddings (E), readout (U), internal blocks (B), and remaining positional/LayerNorm parameters (R). Restoring pre-relapse B raises trough accuracy from 74.76% to 99.85%, whereas restoring $E + U$ reaches 81.45%; injecting trough $E + U$ into the pre-relapse state leaves accuracy at 100%. Thus, on this checkpoint pair, the functional deficit is concentrated primarily in internal blocks.

From the same pre-relapse state and optimizer history, freezing either $E + U$ or B prevents the impending 50-step dip on the probed trajectory. Net internal-block displacement over that interval is predominantly tangential ($\geq 88.4\%$), and applying the measured tangential component reproduces low accuracy whereas the radial component does not. On this trajectory, these interventions implicate joint co-adaptation and tangential internal-block motion in the collapse.

An exploratory, test-conditioned probe analysis examines four qualified relapses selected from 12 trajectories (three Scratch and one $E + U$ transfer; Appendix A). Refitting a linear readout on trough features improves accuracy by 8.37–11.03 points. Trough probes are less accurate than their pre-relapse counterparts at every tested calibration budget; at $N = 2554$, deficits range from 2.66 to 20.00 points. Together, these probes show that trough representations retain recoverable target information but become substantially harder to decode linearly. Cross-initialization differences remain unresolved.

Table 2: **Online Dynamic Validation Gating Across Four Prospective Confirmatory Cohorts ($p = 113$, $H = 1{,}500$ Steps Post-Trigger).** Evaluation of online carrier freezing upon reaching validation accuracy $\geq 95\%$ across verification ($2a+b$, 1-layer, Blocks 250–261, $N = 12$), prospective affine ($3a + b$, Blocks 270–281, $N = 12$), prospective nonlinear quadratic ($a^2 + b^2$, Blocks 290–301, $N = 12$), and prospective 2-layer target ($2a + b$, Blocks 310–321, $N = 12$) cohorts, contrasted with unshielded AdamW and Trigger Snapshot (early stopping). Uncertainties are Mean ± SEM across seeds (p_{flip} from exact two-sided sign-flip permutation tests). By construction, Trigger Snapshot terminates optimization at the trigger, yielding strictly zero post-trigger drawdown. Offline shielding and geometric update ablations are detailed in §5–§6 and the supplementary archive.

Target Cohort / Policy	Trigger t_{trig}	True Max DD (%)	Min Test Acc (%)	Terminal Acc (%)	Prospective Effect (p_{flip})
Cohort 1: Verification Target ($a + b \to 2a + b$, $N = 12$ Blocks 250–261, 1-Layer, $H = 1{,}500$ Steps)					
Trigger Snapshot (Stop at t_{trig})	504 ± 48	$\mathbf{0.00}\%$	$\mathbf{99.21}\% \pm \mathbf{0.14}\%$	$99.21\% \pm 0.14\%$	Early Stopping Baseline (halted at trigger)
$2a + b$ Standard AdamW	504 ± 48	$22.06\% \pm 2.47\%$	$77.88\% \pm 2.47\%$	$\mathbf{99.80}\% \pm \mathbf{0.09}\%$	Unshielded Relapse Baseline
$2a + b$ Online Freeze $E + U$	504 ± 48	$\mathbf{0.60}\% \pm \mathbf{0.17}\%$	$\mathbf{98.94}\% \pm \mathbf{0.20}\%$	$99.36\% \pm 0.11\%$	$+\mathbf{21.45}$ pp reduction, $12/12$ wins ($p = 0.000488^{**}$)
Cohort 2: Prospective Affine Extension ($a + b \to 3a + b$, $N = 12$ Fresh Blocks 270–281, 1-Layer)					
Trigger Snapshot (Stop at t_{trig})	1042 ± 62	$\mathbf{0.00}\%$	$\mathbf{98.86}\% \pm \mathbf{0.24}\%$	$98.86\% \pm 0.24\%$	Early Stopping Baseline (halted at trigger)
$3a + b$ Standard AdamW	1042 ± 62	$28.56\% \pm 2.37\%$	$71.18\% \pm 2.38\%$	$\mathbf{98.89}\% \pm \mathbf{1.10}\%$	Unshielded Relapse Baseline
$3a + b$ Online Freeze $E + U$	1042 ± 62	$\mathbf{5.09}\% \pm \mathbf{1.48}\%$	$\mathbf{94.28}\% \pm \mathbf{1.58}\%$	$97.79\% \pm 0.90\%$	**Affine Confirmed**: $+23.47$ pp, $12/12$ ($p = 0.000488^{**}$)
Cohort 3: Prospective Nonlinear Quadratic Extension ($a + b \to a^2 + b^2$, $N = 12$ Fresh Blocks 290–301, 1-Layer)					
Trigger Snapshot (Stop at t_{trig})	229 ± 7	$\mathbf{0.00}\%$	$\mathbf{99.87}\% \pm \mathbf{0.07}\%$	$99.87\% \pm 0.07\%$	Early Stopping Baseline (halted at trigger)
$a^2 + b^2$ Standard AdamW	229 ± 7	$11.09\% \pm 0.99\%$	$88.91\% \pm 0.99\%$	$\mathbf{99.99}\% \pm \mathbf{0.01}\%$	Unshielded Relapse Baseline
$a^2 + b^2$ Online Freeze $E + U$	229 ± 7	$\mathbf{0.15}\% \pm \mathbf{0.02}\%$	$\mathbf{99.80}\% \pm \mathbf{0.07}\%$	$99.91\% \pm 0.02\%$	**Quadratic Confirmed**: $+10.94$ pp, $12/12$ ($p = 0.000488^{**}$)
Cohort 4: Prospective 2-Layer Target Architecture ($a + b \to 2a + b$, $N = 12$ Fresh Blocks 310–321, 2-Layer)					
Trigger Snapshot (Stop at t_{trig})	1508 ± 255	$\mathbf{0.00}\%$	$\mathbf{97.78}\% \pm \mathbf{0.45}\%$	$97.78\% \pm 0.45\%$	Early Stopping Baseline (halted at trigger)
2-Layer Standard AdamW	1508 ± 255	$80.58\% \pm 1.86\%$	$18.82\% \pm 1.77\%$	$90.23\% \pm 6.30\%$	Unshielded Relapse Baseline
2-Layer Online Freeze $E + U$	1508 ± 255	$\mathbf{65.23}\% \pm \mathbf{2.23}\%$	$\mathbf{33.62}\% \pm \mathbf{2.15}\%$	$\mathbf{98.85}\% \pm \mathbf{0.44}\%$	**Two-Layer Confirmed**: $+15.35$ pp, $11/12$ ($p = 0.00195^{**}$)

5 Targeted Protection: Offline Shielding and Online Dynamic Gating

5.1 Intervention Formulation

Guided by our diagnostic findings, Phase C formulates targeted parameter-level interventions designed to shield transferred computation against optimization collapse: (1) **Carrier Freezing (`carrier_frozen`)**: Representation matrices $\theta \in \{W_E, W_U\}$ are permanently frozen ($\nabla_\theta = 0$, weight decay deactivated: $\theta(t) \equiv \theta(0)$); positional embeddings, LayerNorms, and internal blocks W_B remain fully trainable. Because standard AdamW couples gradient steps and weight decay ($\theta_{t+1} = \theta_t - \eta g_t - \eta\lambda\theta_t$), parameter freezing bundles data-gradient suppression with the arrest of weight decay erosion. Our differential learning rate ablation directly tests whether soft attenuation suffices under the tested $\lambda = 1.0$ schedule. (2) **Differential Learning Rate (`carrier_lr_0.1`)**: Soft dampening with $\eta_{\text{carrier}} = 10^{-4}$ ($0.1\times$ base LR) and base weight decay $\lambda = 1.0$. (3) **Unconfounded Warmup Freezing (`freeze_warmup_500_norest`)**: Representation carriers are frozen for 500 steps, then released without resetting optimizer state.

5.2 Confirmatory Battery Results ($N = 10$ Independent Blocks 230–239, 60 Runs)

We evaluate all interventions across 10 fresh, prospectively specified blocks (60 runs × 10,000 steps). Figure 1 presents the trajectory dynamics and confirmatory results; detailed numerical summaries are reported below and archived in the supplementary materials.

Relapse Suppression via Carrier Freezing. Under standard fine-tuning (`E+U_standard`), 100% (10/10) of runs suffer post-reach relapse, with a mean post-peak dip of 19.40% and mean max drawdown of 20.34%. Permanently freezing the carriers (`E+U_carrier_frozen`) virtually eliminates relapse: post-peak relapse dip is slashed from 19.40% to $\mathbf{0.07}\%$ ($+19.33$ pp reduction, 95% BCa CI: $[14.91, 24.69]$, Holm $p = 0.005859 < 0.01$), and supplementary post-reach max drawdown is slashed from 20.34% to $\mathbf{0.09}\%$ ($+20.26$ pp reduction, 95% BCa CI: $[15.75, 24.96]$, Holm $p = 0.005859 < 0.01$). Across all 10 independent blocks, the contrast is unanimously concordant (10/10 pairs positive), confirming carrier-freezing relapse suppression.

Tested Soft Interventions Fail to Suppress Relapse. Neither $0.1\times$ carrier LR nor 500-step temporary freezing significantly suppresses relapse under the tested schedules (20.80% and 14.12% mean dip, Holm $p = 0.820$ and $p = 0.242$); delayed relapse remains within the 10k-step horizon.

Early Velocity Non-Inferiority Boundary. We tested whether representation-only frozen transfer (`E+U_carrier_frozen`) could match the early learning velocity of full-factor standard transfer (`full_standard`) within a pre-specified non-inferiority margin of $\delta = 100.0$ integral units. The

empirical mean gap is $\mathbf{-144.93}$ **units** (95% BCa CI: $[-164.54, -127.21]$). A one-sided shifted permutation test yields $p = 0.9990 > 0.05$, officially failing non-inferiority. As visible in Figure 1(a) inset, `full_standard` crosses 90% accuracy at step 50, whereas `E+U_carrier_frozen` requires 250 steps. Representation-only shielding did not match the early efficacy of full transfer under the pre-specified non-inferiority criterion.

5.3 Resolving the 4-Quadrant Lifecycle: Takeoff and Stability Synergies

The confirmatory results separate two effects: internal-parameter transfer improves early takeoff, whereas freezing $E + U$ improves stability. In an exploratory 10-run extension, combining full transfer with carrier freezing preserves 50-step takeoff while reducing relapse to 0.02%.

5.4 Online Causal Protection via Dynamic Validation Gating

While offline carrier freezing from initialization achieves stability, it requires deciding parameter shielding ahead of time. To eliminate this constraint, we formulate **Online Dynamic Validation Gating**: models begin with unshielded optimization; upon reaching validation accuracy $\geq 95\%$ for two consecutive evaluations (cadence $\Delta = 50$), representation matrices W_E, W_U are dynamically frozen for subsequent training over a pre-specified finite observation horizon of $H = 1{,}500$ steps post-trigger. We emphasize that our causal stability claim evaluates finite-horizon trajectory robustness ($H = 1{,}500$ steps); long-term stability under unbounded post-trigger optimization remains open. After repairing optimizer-state aliasing in an earlier implementation, we re-evaluated the verification cohort (Blocks 250–261, $N = 12$) with deep-copy isolation, state-hash assertions, and execution-order checks (Table 2, Cohort 1): (1) **Drawdown Suppression**: Under standard training, models suffer severe post-trigger collapse (mean True Max Drawdown 22.06%, min test accuracy 77.88%). Online carrier freezing slashes True Max Drawdown to $\mathbf{0.60}\%$ (median 0.42%, min accuracy 98.94%), yielding an average reduction of $\mathbf{+21.45}$ **percentage points** (95% BCa CI: $[+17.76, +27.06]$, $12/12$ concordant pairs, exact two-sided sign-flip $p = 0.000488 < 0.001$; development cohort replicates with $12/12$ wins, $20.20\% \to 1.83\%$, $p = 0.000488$). (2) **Geometric Constraints**: Pinned direction with free norm (`EU_radial`) yields 37.36% drawdown and is worse than pinned norm with free direction (`EU_tangential`, 5.44%) in all 12 seeds. Complete freezing yields 0.60% drawdown and beats pinned norm in 9/12 seeds ($p = 0.005371$). The permitted update subspaces therefore produce sharply different stability outcomes. (3) **Baseline Triad and Cohort Distinction**: Post-breakthrough degradation admits distinct operational solutions: (i) *Early Stopping* halts at the validation trigger ($t_{\text{trig}} = 504 \pm 48$), yielding $99.21\% \pm 0.14\%$ test accuracy with strictly 0.00% drawdown; (ii) *Retrospective Checkpointing* selects the peak validation model across the trajectory. In offline Phase C (Blocks 230–239), unshielded transfer achieves peak test accuracy of $99.92\% \pm 0.04\%$ (step 910 ± 104) for `E+U_standard` (despite a 19.40% relapse dip) and $99.97\% \pm 0.03\%$ (step 85 ± 8) for `full_standard` (despite a 7.21% dip). Offline retrospective selection and online streaming cohorts represent distinct operational regimes and must not be conflated as direct numerical competitors. Early stopping and retrospective checkpointing address static model delivery; online carrier freezing addresses the distinct regime in which target optimization continues beyond the validation trigger, maintaining higher trajectory mean accuracy (99.55% vs. 99.37%) at a minor terminal cost (-0.44 pp, 99.36% vs. 99.80%).

5.5 Prospective Confirmation Across Tasks and Architecture Depths

Affine Extension ($3a + b \pmod{113}$). On $N = 12$ fresh blocks (Blocks 270–281; all triggered at steps 750–1,350), unshielded fine-tuning undergoes acute relapse (True Max Drawdown 28.56%, range $17.58\% \sim 45.57\%$, min accuracy 71.18%). Online carrier freezing slashes True Max Drawdown to $\mathbf{5.09}\%$ (median 3.55%, min accuracy $\mathbf{94.28}\%$), achieving an average reduction of $\mathbf{+23.47}$ **percentage points** (95% BCa CI: $[+19.56, +29.37]$, $12/12$ concordant pairs, $p = 0.000488 < 0.01$), confirming the primary prospectively specified hypothesis. The secondary boundary prediction ($\leq 2.0\%$ cohort mean drawdown) was not supported (cohort mean 5.09%). Terminal test accuracy under online freeze is $97.79\% \pm 0.90\%$ (-1.10 pp vs. unshielded), confirming that online freezing stabilizes continued downstream optimization.

Nonlinear Quadratic Extension ($a^2 + b^2 \pmod{113}$). Prospective evaluation on the nonlinear quadratic target $a + b \to a^2 + b^2 \pmod{113}$ across $N = 12$ fresh blocks (Blocks 290–301; all triggered at steps 200–250, $t_{\text{trig}} = 229 \pm 7$) confirms that unshielded models experience substantial post-trigger collapse (mean True Max Drawdown $11.09\% \pm 0.99\%$, range $5.83\% \sim 16.33\%$, min accuracy 88.91%). Online carrier freezing slashes True Max Drawdown to $\mathbf{0.15}\% \pm \mathbf{0.02}\%$ (range

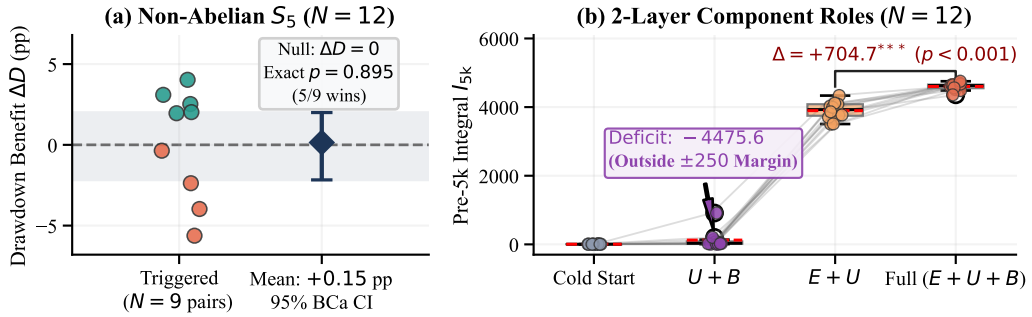

Figure 3: **Generality and Dynamical Boundaries.** (a) Non-abelian S_5 dynamic protection ($N = 12$ prospective blocks, 9 triggered pairs): carrier shielding shows no confirmed drawdown benefit ($\Delta D = +0.15 \pm 1.14$ pp, exact $p = 0.895$). (b) Two-layer component roles (I_{5k}, $N = 12$ fresh blocks): internal blocks accelerate acquisition ($+704.67$ units, $p < 0.001$), while omitting donor embeddings ($U + B$) incurs a -4475.63-unit deficit exceeding the pre-specified ± 250-unit equivalence margin.

$0.01\% \sim 0.30\%$, min accuracy $\mathbf{99.80\% \pm 0.07\%}$), achieving an average reduction of $+\mathbf{10.94}$ **percentage points** (95% BCa CI: $[+9.13, +12.74]$, $12/12$ positive wins, $p = 0.000488 < 0.01$) with minimal terminal cost (-0.09 pp, 99.91% vs. 99.99%). Prospective confirmation extends the protection effect beyond affine targets to a nonlinear quadratic operator ($a^2 + b^2 \pmod{113}$).

Cross-Depth Extension (2-Layer Target Architecture). Prospective evaluation confirms that the protective effect extends from one- to two-layer target models: qualified 1-layer donors ($a + b$) transferred $E + U$ direction to a 2-layer target ($2a + b \pmod{113}$), $L = 2,425{,}472$ parameters, cold internal blocks) across $N = 12$ fresh blocks (Blocks 310–321, triggered at $t_{\text{trig}} = 1508 \pm 255$). Unshielded fine-tuning suffers severe collapse (True Max Drawdown $80.58\% \pm 1.86\%$, min accuracy 18.82%). Carrier shielding prospectively attenuates post-breakthrough drawdown in two-layer targets, reducing True Max Drawdown to $\mathbf{65.23\% \pm 2.23\%}$ ($+\mathbf{15.35}$ **percentage points** reduction, 95% BCa CI: $[+9.30, +21.01]$, $11/12$ concordant wins, $p = 0.001953 < 0.01$). Terminal accuracy was descriptively higher under freezing (98.85% vs. 90.23%), partly because several standard branches failed to recover within the observation window. Carrier shielding reduces drawdown by 15.35 pp at two layers, but the 65.23% residual drawdown shows that protection is partial rather than complete.

6 Generality Boundaries: Non-Abelian Groups and Layer Depth

6.1 Non-Abelian Conjugation: Transient Transfer and Empirical Boundary

To probe boundaries beyond abelian arithmetic, Phase 1 evaluated transfer from $a \cdot b$ to conjugation $a \cdot b \cdot a^{-1}$ on non-abelian S_5 ($|S_5| = 120$, sequence format $[a, b]$, predicting permutation index $c = a \circ b \circ a^{-1}$, 14,400 pairs total partitioned into $N_{\text{train}} = 4{,}320$ [30%], $N_{\text{val}} = 2{,}880$ [20%], $N_{\text{test}} = 7{,}200$ [50%]; 60k steps; Figure 3(a)). Transferred runs surge to peak accuracy of **95.39%** (Block 25) and **84.62%** (Block 31) versus cold start (5.83%), yet fail persistent confirmation ($0/4$ at $q = 0.90$). In a prospective $N = 12$ cohort (Blocks 350–361), triggering required two consecutive validation evaluations ≥ 0.70 at cadence $\Delta = 50$ within $M_{\text{pre}} = 35{,}000$ steps (achieved by 9/12 blocks at $71.22\% \pm 0.51\%$). Over horizon $H = 3{,}000$ steps, carrier shielding showed no confirmed drawdown benefit on triggered pairs ($\Delta D = +\mathbf{0.15} \pm \mathbf{1.14}$ pp, 95% BCa CI: $[-2.17, +2.00]$ pp, exact sign-flip $p = 0.894531$). In the intention-to-treat policy estimand, the 3 untriggered blocks continue unshielded ($\Delta D = 0$), yielding full-cohort policy effect $\Delta D = +0.11 \pm 0.84$ pp (95% BCa CI: $[-1.72, +1.47]$ pp), providing a cross-family boundary case.

6.2 Two-Layer Component Roles: Acquisition and Embeddings

Following exploratory runs right-censored at 30k steps ($p = 0.2500$), we evaluated a 60-run confirmatory battery across $N = 12$ fresh blocks (Blocks 330–341, 10k steps; comprising 12 donor training runs on 2-layer $a + b$ plus 48 target runs across 4 conditions: ‘cold‘, ‘full‘, ‘E+U‘, and ‘U+B‘; Figure 3(b)). Donors are 2-layer Transformers ($L = 2{,}425{,}472$ parameters); transfer maps

internal blocks layer-by-layer ($B_0 \to B_0, B_1 \to B_1$) alongside representation matrices (W_E, W_U), with cold Frobenius scale matching. (1) **Internal Acceleration**: Internal blocks accelerate acquisition beyond $E + U$ in two-layer models (12/12 wins, $I_{5k} = 4600.9 \pm 31.3$ vs 3896.3 ± 73.4, gain $+\mathbf{704.67}$ units, 95% BCa CI: $[+549.86, +870.00]$, exact $p = 0.000488 < 0.001$; trigger onset $1{,}283 \to 158$ steps). (2) **Embedding Requirement**: Transferring $U + B$ without donor embeddings fails to preserve Full-transfer acquisition ($I_{5k} = 125.3 \pm 73.7$, deficit $-\mathbf{4475.63}$ units, 95% BCa CI: $[-4570.56, -4229.47]$; untriggered in 10/12 blocks), falling far outside the ± 250-unit equivalence margin (TOST $p = 1.000000$). Donor embeddings are thus required to preserve Full-transfer acquisition in two-layer models.

7 Related Work

Delayed generalization in modular arithmetic (Power et al., 2022; Liu et al., 2022) reflects trigonometric circuits (Nanda et al., 2023; Varma et al., 2023) and lazy-to-rich transitions (Kumar et al., 2024); scale matching isolates directional transfer (Hagmann et al., 2023). While Xu et al. (2025) show that pre-trained embeddings accelerate single-task grokking, our factorial decomposition separates internal computation from representation carriers: internal blocks accelerate acquisition with transferred carriers, while donor embeddings are required in two-layer models. Closest to our stability analysis, Janati et al. (2026) show that parameter freezing prevents circuit unlearning in single-task grokking. We investigate cross-operator transfer, establish a dissociation between acquisition velocity and retention stability, and introduce dynamic validation gating to stabilize continual post-trigger optimization.

8 Discussion, Limitations, and Conclusion

This work characterizes the algorithmic transfer lifecycle in grokking: (1) **Acquisition**: Internal-block transfer accelerates acquisition beyond carriers, replicating at two layers ($+704.67$ units, $p < 0.001$). Co-transferring internal blocks and heads satisfies the pre-specified mean-latency equivalence criterion to Full in one-layer models, whereas donor embeddings are required in two-layer models ($U + B$ falls 4475.6 units below Full, TOST $p = 1.00$). (2) **Relapse**: Factorial state replacement localizes post-transfer deficits to internal blocks; linear readout refitting recovers 8.37–11.03 points without parameter retraining. (3) **Protection and Boundaries**: Online carrier shielding suppresses acute relapse in one-layer tasks ($+21.45$, $+23.47$, $+10.94$ pp) and attenuates drawdown in two-layer targets ($80.58\% \to 65.23\%$, $+15.35$ pp), though residual instability remains at two layers. While early stopping secures static accuracy ($\geq 98.8\%$), online carrier freezing stabilizes continual optimization beyond the trigger. In contrast, prospective S_5 transfer did not reproduce shielding benefits ($\Delta D = +0.15 \pm 1.14$ pp, exact $p = 0.895$), providing a cross-family boundary case. Whether component specialization generalizes to deeper architectures and non-algorithmic domains remains open. Together, these findings support a component-level account of transfer acceleration and continual stability in algorithmic Transformers.

Reproducibility Statement. Architectures, splits, hyperparameters, protocols, and statistical tests are in the text and appendices; run artifacts contain deterministic seeds, checkpoints, and reproduction scripts. **AI Use Statement.** AI tools assisted with code development and manuscript editing. The authors designed the study, audited numerical claims against saved artifacts, and remain responsible for the scientific content.

References

Andrey Gromov. Grokking modular arithmetic. *arXiv preprint arXiv:2301.02679*, 2023. URL https://arxiv.org/abs/2301.02679v1.

Michael Hagmann, Philipp Meier, and Stefan Riezler. Towards inferential reproducibility of machine learning research. *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2302.04054v7.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. URL https://arxiv.org/abs/1606.08415.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Ali Janati, Kaoutar El Maghraoui, Andrei Kanavalau, and Anass Belfatmi. Post-grokking collapse at the representation-readout interface in muon-trained transformers. *arXiv preprint arXiv:2608.07436*, 2026. URL https://arxiv.org/abs/2608.07436.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations (ICLR)*, 2022.

Tanishq Kumar, Blake Bordelon, Samuel J. Gershman, and Cengiz Pehlevan. Grokking as the transition from lazy to rich training dynamics. *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2310.06110v3.

Ziming Liu, Ouail Kitouni, Niklas Nolte, Eric J. Michaud, Max Tegmark, and Mike Williams. Towards understanding grokking: An effective theory of representation learning. *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2205.10343v2.

Ziming Liu, Eric J. Michaud, and Max Tegmark. Omnigrok: Grokking beyond algorithmic data. *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2210.01117v2.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *International Conference on Learning Representations (ICLR)*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Mohamad Amin Mohamadi, Zhiyuan Li, Lei Wu, and Danica J. Sutherland. Why do you grok? a theoretical analysis of grokking modular addition. *International Conference on Machine Learning (ICML)*, 2024. URL https://arxiv.org/abs/2407.12332v1.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2301.05217v3.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022. URL https://arxiv.org/abs/2201.02177v1.

Donald J. Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6):657–680, 1987.

Vimal Thilak, Etai Littwin, Shuangfei Zhai, Omid Saremi, Roni Paiss, and Joshua Susskind. The slingshot mechanism: An empirical study of adaptive optimizers and the grokking phenomenon. *arXiv preprint arXiv:2206.04817*, 2022.

Vikrant Varma, Rohin Shah, Zachary Kenton, Janós Kramár, and Ramana Kumar. Explaining grokking through circuit efficiency. *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2309.02390v1.

Zhiwei Xu, Zhiyu Ni, Yixin Wang, and Wei Hu. Let me grok for you: Accelerating grokking via embedding transfer from a weaker model. *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/fa5ddd6bac0d665c72969d79221b680a-Abstract-Conference.html.

A Diagnostic Linear Probing Specification, Optimization Convergence Checks, and Full Accounting

This appendix provides the full technical specification, optimization convergence checks, trajectory event accounting, and numerical results for the diagnostic linear probing investigation.

A.1 Diagnostic Figures

Grassmann Subspace Alignment Metric. In Figure 4(b), the alignment between the token embedding representation $W_E(t) \in \mathbb{R}^{p \times d_{\text{model}}}$ at step t and the donor initialization $W_E(0)$ is quantified via the mean canonical principal angle cosine across their leading $k = 12$ left singular subspaces:

$$\cos \bar{\theta}(t) = \frac{1}{k} \sum_{i=1}^{k} \sigma_i(Q(t)^\top Q^{\text{donor}}) = \frac{1}{k} \sum_{i=1}^{k} \cos \theta_i, \tag{4}$$

where $Q(t) \in \mathbb{R}^{p \times k}$ and $Q^{\text{donor}} \in \mathbb{R}^{p \times k}$ are orthonormal bases spanning the top-k left singular vectors of $W_E(t)$ and $W_E(0)$ respectively, and σ_i are the singular values of their inner projection matrix.

A.2 Protocol Specification

Motivation and Scope. Following recent analyses distinguishing circuit failure from circuit masking (Janati et al., 2026), we investigate whether post-grokking relapse reflects an intrinsic loss of decodable target information within internal representations or a geometric misalignment between internal features and the readout classification head. We evaluate an exploratory development cohort consisting of 12 complete training trajectories across Blocks 210–213 spanning three initialization conditions:

1. `scratch`: standard cold initialization (W_b^{cold});
2. `transfer_EU`: token embeddings W_E and classification head W_U initialized from trained donor checkpoints with per-tensor Frobenius scale matching ($\theta^{\text{transfer}} = (\theta^{\text{donor}}/(\|\theta^{\text{donor}}\|_F + 10^{-12})) \cdot \|\theta^{\text{cold}}\|_F$), with internal transformer blocks and LayerNorms initialized cold;
3. `transfer_full`: all 2D weight matrices (W_E, W_U, and internal self-attention and MLP weights W_B) transferred with per-tensor scale matching, while 1D LayerNorm parameters and positional embeddings remain cold.

Representation Extraction and Consistency Verification. Representations $z \in \mathbb{R}^{d_{\text{model}}}$ ($d_{\text{model}} = 128$) are extracted at the sequence-level token position $z = h_L[:, -1, :]$ immediately following final LayerNorm, prior to the readout matrix $W_U \in \mathbb{R}^{p \times d_{\text{model}}}$. For every evaluated checkpoint, we computationally asserted that applying the existing readout head to z exactly reproduces the forward model logits:

$$\max_{x \in \mathcal{D}_{\text{test}}} \|W_U z(x) - f(x)\|_\infty < 10^{-5}, \tag{5}$$

guaranteeing that probed representations correspond precisely to the operational input to the classification head.

Event Detection via Running Peak Drawdown. To avoid censoring early relapses occurring prior to the global trajectory maximum, events are tracked via running peak drawdown. Denoting running peak accuracy by $M(t) = \max_{s \le t} \text{Acc}(s)$ after validation breakthrough (ValAcc $\ge 95\%$ for two consecutive checks at cadence $\Delta = 50$):

1. **Pre-Relapse Peak** (t_{pre}): the running peak immediately preceding a drop;
2. **Relapse Trough** (t_{trough}): the local minimum where drawdown $\Delta_{\text{dip}}(t) = M(t) - \text{Acc}(t) \ge 10.0$ pp and $\text{Acc}(t) < 90.0\%$;
3. **Post-Recovery** (t_{recov}): the first subsequent evaluation where $\text{Acc}(t) \ge 95.0\%$.

Events are searched within 2,000 steps after breakthrough. Scratch runs have a 30,000-step budget and transfer runs a 5,000-step budget. Runs without a qualifying drop are labeled `CENSORED_NO_RELAPSE`; runs without breakthrough are `CENSORED_UNBROKEN`. If a drop occurs without recovery, its Pre–Trough probes remain in the analysis and the recovery value is censored.

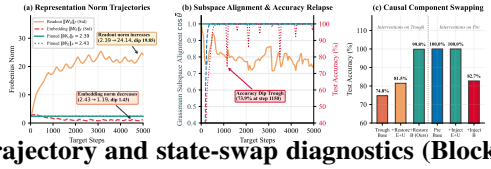

Figure 4: **Exploratory trajectory and state-swap diagnostics (Blocks 210/230).** (a–b) Embedding/readout norms and embedding-subspace alignment evolve during unshielded training; alignment is not monotone in accuracy. (c) On one pre-relapse/trough checkpoint pair, restoring internal blocks recovers 99.85% accuracy, whereas restoring $E + U$ reaches 81.45%. These single-trajectory observations localize the contemporaneous functional deficit but do not identify a general cause.

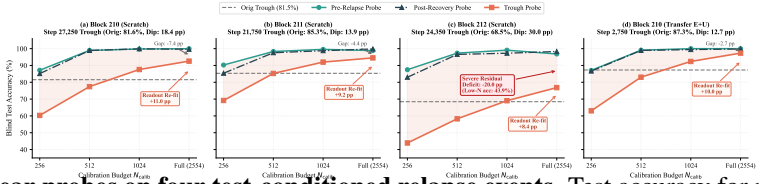

Figure 5: **Linear probes on four test-conditioned relapse events.** Test accuracy for probes fitted to pre-relapse, trough, and recovery representations across calibration budgets. Trough probes improve with more data but remain below paired pre-relapse probes under the tested protocol. This finite-budget deficit neither establishes irreversible information loss nor excludes a better decoder.

Linear Probe Objective and Subsetting. At each checkpoint stage, an independent linear probe $U \in \mathbb{R}^{p \times d_{\text{model}}}$ is fitted by minimizing regularized empirical cross-entropy over a calibration subset $\mathcal{S}_N \subset \mathcal{D}_{\text{train}}$:

$$\min_{U \in \mathbb{R}^{p \times d_{\text{model}}}} \mathcal{L}(U) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{\text{CE}}(U z_i, y_i) + \frac{\lambda}{2} \|U\|_F^2, \tag{6}$$

across sample budgets $N \in \{256, 512, 1024, 2554\}$, where $N = 2554$ is the full training partition. Subsets $\mathcal{S}_{256} \subset \mathcal{S}_{512} \subset \mathcal{S}_{1024} \subset \mathcal{S}_{2554}$ are constructed as strictly nested leading prefixes ($\mathcal{D}_{\text{train}}[:N]$) under the canonical deterministic train/val/test split (seed `block` + 1000). Candidate penalties $\lambda \in \{0, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$ are selected by validation cross-entropy and the selected probe is evaluated on the test partition. Because event times are selected using test accuracy, this is a test-conditioned diagnostic rather than a globally blind test.

A.3 Optimization Convergence Checks

Random Stream Isolation. To prevent pseudo-random coupling between trajectory execution and probe fitting, each training run used a dedicated PyTorch CUDA generator (`torch.Generator(device='cuda')`) for mini-batch sequencing. Probe initialization weights were generated using an independent CPU generator initialized with a fixed seed (`seed=42`).

Solver and Gate. The 288 candidate fits use L-BFGS with strong Wolfe line search (maximum 500 iterations). Of these, 286 satisfy the pre-specified gate $\|\nabla \mathcal{L}\|_F < 10^{-3}$; the two exceptions occur at $\lambda = 1$, $N = 2554$ and are not selected. All 48 validation-selected probes pass the gate. A separate Block-210 implementation check compared Adam and L-BFGS at one trough configuration and found a 0.03-point test-accuracy difference; it was not a second solver applied to all fits. For $\lambda = 0$, we report a finite-budget unregularized solution and do not claim a unique finite optimum.

A.4 Full Trajectory Event Accounting

Table 3 reports the complete trajectory event accounting across all 12 runs. Notably, all 3 broken from-scratch trajectories underwent deep qualified relapses (dips of 18.37, 13.86, and 29.95 pp), disproving the assumption that unshielded training without transfer converges smoothly.

A.5 Full Linear Probe Decodability and Paired Deficits

Table 4 presents test accuracies across all calibration budgets $N \in \{256, 512, 1024, 2554\}$ for the 4 qualified events.

Table 3: **Event Accounting and Trajectory Census Across 12 Runs (Blocks 210–213).** Breakthrough is defined as the first step with ValAcc $\geq 95\%$ for two consecutive evaluations. Pre-relapse peak, relapse trough, and recovery steps are detected via running peak drawdown. All 3 broken Scratch runs suffered deep qualified relapses (≥ 10 pp dip and test accuracy $< 90\%$), showing that scratch relapse occurs in this development cohort.

Run ID	Condition	Breakthrough	Pre-Peak (t_{pre})	Trough (t_{trough})	Trough Acc	Drawdown Dip	Recovery (t_{recov})	Accounting Status
b210_scratch	Scratch	Step 26,600	Step 27,100 (99.92%)	Step 27,250	81.55%	18.37 pp	Step 27,300 (99.71%)	**QUALIFIED_EVENT**
b211_scratch	Scratch	Step 21,300	Step 21,550 (99.19%)	Step 21,750	85.33%	13.86 pp	Step 21,800 (99.07%)	**QUALIFIED_EVENT**
b212_scratch	Scratch	Step 23,850	Step 24,150 (98.41%)	Step 24,350	68.46%	29.95 pp	Step 24,400 (97.61%)	**QUALIFIED_EVENT**
b213_scratch	Scratch	Unbroken (> 30k)	—	—	—	—	—	CENSORED_UNBROKEN
b210_transfer_EU	Transfer E+U	Step 850	Step 1,200 (100.00%)	Step 2,750	87.30%	12.70 pp	Step 2,800 (100.00%)	**QUALIFIED_EVENT**
b211_transfer_EU	Transfer E+U	Step 850	—	—	—	6.42 pp	—	CENSORED_NO_RELAPSE
b212_transfer_EU	Transfer E+U	Step 500	—	—	—	9.57 pp	—	CENSORED_NO_RELAPSE
b213_transfer_EU	Transfer E+U	Step 500	—	—	—	1.20 pp	—	CENSORED_NO_RELAPSE
b210_transfer_full	Transfer Full	Step 100	—	—	—	0.94 pp	—	CENSORED_NO_RELAPSE
b211_transfer_full	Transfer Full	Step 100	—	—	—	5.43 pp	—	CENSORED_NO_RELAPSE
b212_transfer_full	Transfer Full	Step 100	—	—	—	4.52 pp	—	CENSORED_NO_RELAPSE
b213_transfer_full	Transfer Full	Step 100	—	—	—	6.89 pp	—	CENSORED_NO_RELAPSE

Table 4: **Linear Probe Decodability Across Checkpoint Stages and Calibration Budgets.** Test accuracy (%) of probes trained on frozen penultimate representations across four test-conditioned relapse events. The penalty λ is selected by validation loss. Paired trough deficit is $\text{Acc}_{\text{trough}}(N) - \text{Acc}_{\text{pre}}(N)$; re-fit gain compares the full-sample trough probe with the unadapted trough model.

Run ID	Stage	Original Acc	$N = 256$	$N = 512$	$N = 1024$	$N = 2554$ **(Full)**	Re-fit Gain	Paired Deficit $\Delta(2554)$
b210_scratch	Pre ($t = 27{,}100$)	99.92%	87.17%	98.96%	99.79%	99.98%	—	—
	Trough ($t = 27{,}250$)	81.55%	60.30%	77.41%	87.59%	92.58%	**+11.03 pp**	**-7.40 pp**
	Recov ($t = 27{,}300$)	99.71%	85.15%	98.90%	99.89%	99.62%	—	—
b211_scratch	Pre ($t = 21{,}550$)	99.19%	90.26%	98.31%	99.57%	98.85%	—	—
	Trough ($t = 21{,}750$)	85.33%	69.18%	85.25%	92.00%	94.50%	**+9.17 pp**	**-4.35 pp**
	Recov ($t = 21{,}800$)	99.07%	85.37%	97.57%	98.78%	99.80%	—	—
b212_scratch	Pre ($t = 24{,}150$)	98.41%	87.48%	97.34%	99.10%	96.83%	—	—
	Trough ($t = 24{,}350$)	68.46%	43.90%	58.26%	69.03%	76.83%	**+8.37 pp**	**-20.00 pp**
	Recov ($t = 24{,}400$)	97.61%	82.98%	96.59%	97.28%	98.30%	—	—
b210_transfer_EU	Pre ($t = 1{,}200$)	100.00%	87.07%	99.07%	100.00%	100.00%	—	—
	Trough ($t = 2{,}750$)	87.30%	63.01%	83.04%	92.39%	97.34%	**+10.04 pp**	**-2.66 pp**
	Recov ($t = 2{,}800$)	100.00%	86.66%	98.95%	99.45%	99.81%	—	—
Summary Statistics: Paired Trough Deficit $\Delta(N) = \text{Acc}_{\text{trough}}(N) - \text{Acc}_{\text{pre}}(N)$								
Scratch ($N = 3$, Sample Std, $ddof = 1$)		—	$-30.51\% \pm 11.68\%$	$-24.56\% \pm 13.27\%$	$-16.61\% \pm 11.88\%$	$-10.58\% \pm 8.30\%$	—	—
Transfer E+U ($N = 1$)		—	-24.05%	-16.03%	-7.61%	-2.66%	—	—

Table 5: **Complete Parameter Census of `GrokTransformer` ($d_{\text{model}} = 128, n_{\text{head}} = 4, d_{\text{mlp}} = 512, p = 113$).** Tensor keys match the exact state dictionary entries in `src/model.py`. Component assignments correspond to the factorial decomposition (E: Token Embedding, U: Readout Head, B: Internal Blocks). Transferred 2D weight matrices are rescaled using per-tensor Frobenius norm matching against the target seed's cold initialization. Positional embeddings, LayerNorm parameters, and MLP biases strictly retain their cold-initialized values in all conditions.

Parameter Key	PyTorch Module	Shape	Parameters	Initialization Scheme	Full Transfer (`full`)
token_emb.weight	nn.Embedding(115, 128)	[115, 128]	14,720	$\mathcal{N}(0, 0.02^2)$	Transferred (E, Frobenius rescaled)
pos_emb.weight	nn.Embedding(4, 128)	[4, 128]	512	$\mathcal{N}(0, 0.02^2)$	**Cold Retained** (indices 0, 1 used)
blocks.0.ln1.weight	nn.LayerNorm(128)	[128]	128	Constant **1**	**Cold Retained**
blocks.0.ln1.bias	nn.LayerNorm(128)	[128]	128	Constant **0**	**Cold Retained**
blocks.0.attn.qkv.weight	nn.Linear(128, 384, bias=False)	[384, 128]	49,152	$\mathcal{N}(0, 0.02^2)$	Transferred (B, Frobenius rescaled)
blocks.0.attn.out.weight	nn.Linear(128, 128, bias=False)	[128, 128]	16,384	$\mathcal{N}(0, 0.02^2)$	Transferred (B, Frobenius rescaled)
blocks.0.ln2.weight	nn.LayerNorm(128)	[128]	128	Constant **1**	**Cold Retained**
blocks.0.ln2.bias	nn.LayerNorm(128)	[128]	128	Constant **0**	**Cold Retained**
blocks.0.mlp.0.weight	nn.Linear(128, 512)	[512, 128]	65,536	$\mathcal{N}(0, 0.02^2)$	Transferred (B, Frobenius rescaled)
blocks.0.mlp.0.bias	nn.Linear(128, 512)	[512]	512	Constant **0**	**Cold Retained**
blocks.0.mlp.3.weight	nn.Linear(512, 128)	[128, 512]	65,536	$\mathcal{N}(0, 0.02^2)$	Transferred (B, Frobenius rescaled)
blocks.0.mlp.3.bias	nn.Linear(512, 128)	[128]	128	Constant **0**	**Cold Retained**
ln_f.weight	nn.LayerNorm(128)	[128]	128	Constant **1**	**Cold Retained**
ln_f.bias	nn.LayerNorm(128)	[128]	128	Constant **0**	**Cold Retained**
head.weight	nn.Linear(128, 113, bias=False)	[113, 128]	14,464	$\mathcal{N}(0, 0.02^2)$	Transferred (U, Frobenius rescaled)
Total Parameters			**227,712** (226,304 in 2D matrices; 1,408 in 1D vectors)		

B Model Architecture, Parameter Keys, and Initialization Details

This appendix provides the complete architectural parameter census, exact PyTorch state-dictionary keys, tensor dimensions, initialization distributions, and parameter transfer conventions for the `GrokTransformer` model defined in `src/model.py` and evaluated across all experiments.

B.1 Architectural Census and Tensor Inventory

Table 5 details the parameter tensors of the 1-layer Transformer architecture. The model is a pre-LayerNorm Transformer with hidden dimension $d_{\text{model}} = 128$, $n_{\text{head}} = 4$ self-attention heads ($d_k = d_v = 32$), and a 2-layer MLP with hidden dimension $d_{\text{mlp}} = 512$ and GELU activations (Hendrycks & Gimpel, 2016). The total parameter count is 227,712.

Two-Layer Architecture Census and Mapping. In the prospective cross-depth and 2-layer role experiments (§5.5, §6), the architecture instantiates two identical transformer blocks (`blocks.0` and `blocks.1`), each containing 197,760 parameters (196,608 in 2D weight matrices; 1,152 in 1D LayerNorm and bias vectors). Together with token embeddings (14,720), positional embeddings (512), final LayerNorm (256), and linear readout head (14,464), the complete 2-layer model contains **425,472** parameters (422,912 in 2D matrices, of which 422,400 are transferable 2D weights; 2,560 in 1D vectors). When transferring internal blocks (B), 2D weight matrices are transferred isomorphically layer-by-layer ($B_0 \to B_0, B_1 \to B_1$), with each tensor rescaled using per-tensor cold Frobenius norm matching. All 1D parameters (LayerNorm weights/biases and MLP biases in both blocks) strictly retain native cold initialization.

B.2 Exact Module Conventions and Transfer Rules

Fused Attention Projections. Self-attention projections are parameterized as a single fused linear layer `blocks.0.attn.qkv.weight` without bias, projecting directly from $d_{\text{model}} \to 3 \times d_{\text{model}}$. Query, key, and value representations are split along the output dimension. The projection is followed by the unbiased linear output projection `blocks.0.attn.out.weight`. In factorial conditions transferring internal blocks (B, $E + B$, $U + B$, and `full`), both matrices are transferred from the qualified donor.

LayerNorm and Bias Vectors Retain Cold Initialization. The architecture uses pre-LayerNorm parameterization: `ln1` normalizes attention inputs, `ln2` normalizes MLP inputs, and `ln_f` normalizes the final sequence representation prior to the linear readout head. All LayerNorm weights and biases are initialized to standard constants (**1** and **0** respectively). Across the three LayerNorm modules, these comprise six one-dimensional vectors (three weights and three biases). MLP biases (`blocks.0.mlp.0.bias` and `blocks.0.mlp.3.bias`) are initialized to **0**, adding two further one-dimensional vectors. Thus the eight one-dimensional vectors together contain 1,408 parameters. Crucially, all 1D parameters are strictly excluded from transfer across all conditions: they retain their native cold initialization.

Positional Embeddings and Vocabulary Structure. The embedding table `token_emb.weight` has vocabulary size $V = p + 2 = 115$ for prime $p = 113$, reserving indices for operands and structural delimiters. The positional embedding table `pos_emb.weight` has capacity $4 \times d_{\text{model}}$; inputs $[a, b]$ index positions 0 and 1. Positional embeddings are not transferred in any condition and retain cold Gaussian initialization.

Per-Tensor Frobenius Norm Scale Matching. Under any transfer condition, each transferred 2D parameter tensor θ inherits the direction of the qualified donor tensor θ^{donor} while matching the Frobenius norm of the target seed's cold initialization:

$$\theta^{\text{transfer}} = \frac{\theta^{\text{donor}}}{\|\theta^{\text{donor}}\|_F + 10^{-12}} \cdot \|\theta^{\text{cold}}\|_F. \tag{7}$$

In the control condition `embed_orig_scale`, `token_emb.weight` is copied directly at the donor's unnormalized scale ($\theta^{\text{transfer}} = \theta^{\text{donor}}$).