IS GROKKING OBSERVABLE FROM REPRESENTATION DYNAMICS

ABSTRACT

Prior work has linked grokking to the evolution of hidden representation geometry, yet it remains unclear whether representation geometry can provide a reliable observable of grokking during learning. To address this question, we introduce M_{int}, a label-free statistic that quantifies pairwise *representation overlap*, *i.e.*, the overlap of hidden features, and identify *geometry-to-behaviour assumptions* under which representation overlap serves as a reliable observable of grokking. Although these assumptions need not hold in general, M_{int} nevertheless captures informative changes in representation overlap throughout training. We further investigate M_{int}-penalized training as a means of reducing representation overlap. Our experiments show that, penalizing the loss by M_{int} within a small range can advance the onset of generalization in practice. Together, our results clarify the conditions under which representation overlap can serve as an observable of grokking. We also identify representation overlap as a practical target for influencing generalization onset, and establish M_{int} as an informative descriptor of representation dynamics throughout learning.

1 INTRODUCTION

Grokking is a delayed transition from training-set fit to generalization, which has become a controlled laboratory for studying how generalization arises from representation dynamics (Power et al., 2022). Mechanistic analyses reveal that grokking can be explained through features (Nanda et al., 2023; Chughtai et al., 2023; Kumar et al., 2024; Varma et al., 2023; Lyu et al., 2024) or circuits (Nanda et al., 2023; Chughtai et al., 2023; Liu et al., 2022; Varma et al., 2023; Lee et al., 2024; Prieto et al., 2025; Thilak et al., 2024). More recent studies track the relationship between feature dynamics and modular tasks, linking grokking to hidden representation geometry (Raghu et al., 2017; Kornblith et al., 2019; Roy & Vetterli, 2007; Levi et al., 2024). Across these studies, a recurring observation emerges: once training accuracy reaches a near-perfect level, hidden representations continue to change substantially rather than converge, more closely mirroring test accuracy, which remains at chance level, than the already-saturated training accuracy (Tian, 2026; He et al., 2026).

This motivates a key question: *can representation geometry provide a reliable observable of grokking computed during learning* that indicates whether and when target-relevant generalization emerges? Recent work has explored geometric measures for early detection of grokking (Golwala, 2026) and shown that geometry-based interventions can influence its onset (Kazanskii, 2026; Howe, 2026). These findings highlight the potential of representation geometry for observing grokking in specific settings, yet leave open the general conditions under which geometric statistics measured during learning can reliably indicate whether and when target-relevant generalization emerges.

To study this question, we introduce M_{int}, a label-free statistic that quantifies pairwise *representation overlap*, *i.e.*, the overlap among hidden features. Specifically, M_{int} measures the root-mean-square off-diagonal cosine overlap among hidden features. In parallel, we quantify target relevance using *target-subspace alignment* $\Omega_y(F)$, defined as the fraction of a fixed target direction y captured by the current representation subspace. Within this $\Omega_y(F)$–M_{int} framework, we identify *geometry-to-behaviour assumptions* under which representation overlap serves as a reliable observable of grokking, and prove a *dynamic unobservability theorem* that establishes the theoretical guarantee. Importantly, although this does not establish M_{int} as an observable of grokking, we prove that M_{int} *co-varies with generalization behaviour* rather than with changes in task-relevant prediction. This establishes M_{int} as a *task-independent geometric statistic* of representation dynamics.

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.

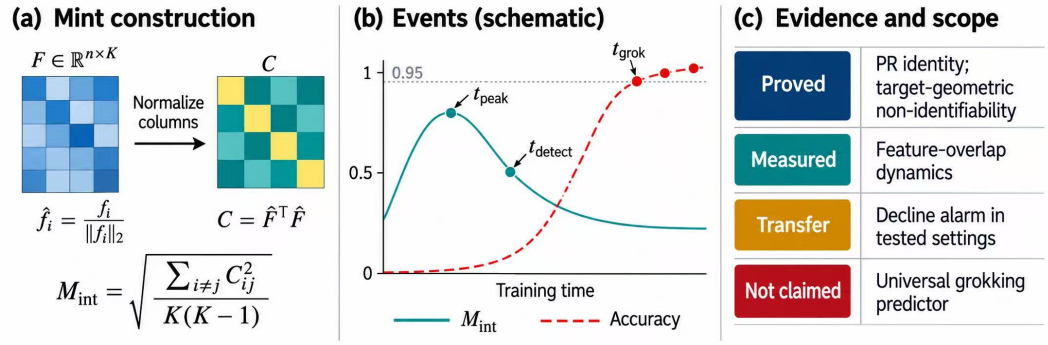

Figure 1: This paper constructs M_{int} (a) and analyzes its relationship with target accuracy (b) to determine when representation geometry provides a reliable observable of grokking. We conclude that the full representation geometry determines the target-geometric transition, whereas any feature-overlap statistic—including the complete history of M_{int}—does not.

We conduct extensive experiments with frozen-hidden, permuted-target, and low-data controls across modular tasks and model circuits to empirically examine these theoretical findings. Although the geometry-to-behavior assumptions need not hold in general, M_{int} nevertheless captures informative changes in representation overlap throughout training. We further investigate M_{int}-penalized training as a means of reducing representation overlap by varying the strength of the M_{int} penalty. The results show that penalizing the loss with M_{int} within a small range can advance the onset of generalization in practice. Together, our results clarify the conditions under which representation overlap can serve as an observable of grokking. We also identify representation overlap as a practical target for influencing generalization onset, and establish M_{int} as an informative descriptor of representation dynamics throughout learning.

Our contributions are:

- We establish a theoretical framework, *e.g.*, $\Omega_y(F)$–M_{int}, connecting representation overlap to generalization behaviour and characterize the geometry-to-behaviour assumptions under which geometric statistics can serve as observables of grokking.
- We introduce M_{int}, a label-free statistic that can capture generalization-related geometry without being dominated by task-relevant prediction.
- We conduct extensive experiments to validate our theoretical findings, showing that representation overlap is a practical target for intervening in the onset of generalization and that M_{int} is an informative descriptor of representation dynamics throughout learning.

2 Related Work

Grokking mechanisms and progress measures. Grokking motivates the search for internal learning progress before test accuracy improves (Power et al., 2022). Existing studies explain delayed generalization through learned features, circuits, and implicit bias (Nanda et al., 2023; Chughtai et al., 2023; Liu et al., 2022; Varma et al., 2023; Lyu et al., 2024; Kumar et al., 2024), with recent work examining feature emergence and modular-addition dynamics (Tian, 2026; He et al., 2026). Following this direction, we study pairwise representation overlap and characterize what its dynamics reveal about learning, including the limits of using overlap to observe grokking.

Representation diagnostics and redundancy. Studying representation changes during grokking requires choosing which geometric properties to measure. SVCCA and CKA compare representations across models or checkpoints (Raghu et al., 2017; Kornblith et al., 2019), while effective rank summarizes dimensionality (Roy & Vetterli, 2007). DeCov, Barlow Twins, and VICReg use feature relationships to reduce redundancy (Cogswell et al., 2016; Zbontar et al., 2021; Bardes et al., 2022). Building on these ideas, we introduce M_{int} to track pairwise cosine overlap without labels, establish its exact relationship to participation-ratio rank, and distinguish raw from centred overlap.

Detection and geometric intervention. Representation changes during grokking raise two practical questions: whether they can signal generalization onset and whether modifying them can make generalization occur earlier. Classical methods such as CUSUM and Bayesian online detection identify changes in time series (Page, 1954; Adams & MacKay, 2007). Recent grokking studies investigate geometric detectors, spectral signals, and circuit synchronisation (Golwala, 2026; Truong et al., 2026; Sivasankar, 2026), as well as geometric controls and task-structured priors for influencing onset (Kazanskii, 2026; Howe, 2026). A recent audit also distinguishes initial onset from long-term convergence (Khanh, 2026). Following these monitoring and intervention directions, we use a simple fixed rule to detect decreases in M_{int} and test penalties that encourage or suppress overlap, with controls that match gradient norms. Our contribution is to examine both their usefulness and their limits: alarms can occur without generalization, and earlier onset can come with poorer retention. Theorem 1 further establishes that feature-overlap histories alone cannot universally determine the target-geometric transition defined in our framework.

3 Representation Geometry Dynamics and Observability

This section places the M_{int} statistic within an observability framework. Section 3.1 defines M_{int} and its exact algebraic structure; Section 3.2 defines the trajectory space and the two geometry observations; Section 3.3 introduces target-subspace alignment and the persistent transition event; Section 3.4 defines observability and its ambiguity quantifier; Section 3.5 states the dynamic unobservability theorem and its consequences; and Section 3.6 states the geometry-to-behaviour assumptions under which the geometric event coincides with behavioural grokking. All proofs are collected in Appendix A. We formalise the information retained by representation overlap and its limits for identifying target-geometric transitions; all proofs appear in Appendix A.

3.1 The Mint statistic

Definition 1 (M_{int})**.** Let $F = [f_1, \dots, f_K] \in \mathbb{R}^{n\times K}$ contain $K \geq 2$ nonzero columns evaluated on n probe examples,and let $G = F^\top F$. define

$$C_{jk} = \frac{G_{jk}}{\sqrt{G_{jj}\,G_{kk}}}, \qquad M_{\mathrm{int}}(F) = \left[\frac{1}{K(K-1)}\sum_{j\neq k} C_{jk}^2\right]^{1/2}. \tag{1}$$

Here C_{jk} is the cosine similarity between features j and k, so M_{int} measures their root-mean-square pairwise overlap. It lies in $[0, 1]$ and is zero exactly when the feature columns are pairwise orthogonal.

Exact structure. For every retained nonzero feature set, $0 \leq M_{\mathrm{int}} \leq 1$, and the quantity is invariant to column permutations and nonzero feature-wise rescaling. It vanishes exactly for pairwise orthogonal columns. Let $U = [f_1/\|f_1\|_2, \dots, f_K/\|f_K\|_2]$, so $C = U^\top U$. Its participation-ratio rank is $r_{\mathrm{PR}}(U) = (\operatorname{tr} C)^2/\|C\|_F^2$. Since $\operatorname{tr} C = K$ and $\|C\|_F^2 = K + K(K-1)M_{\mathrm{int}}^2$, we obtain

$$r_{\mathrm{PR}}(U) = \frac{K}{1 + (K-1)M_{\mathrm{int}}^2}, \qquad M_{\mathrm{int}}^2 = \frac{K/r_{\mathrm{PR}} - 1}{K-1}. \tag{2}$$

For fixed K, lower overlap therefore corresponds to higher participation-ratio rank. When the rank of the matrix, $\operatorname{rank}(U) < K$, overlap has a positive lower bound; the exact bound and other basic properties appear in Appendix A.1. Raw M_{int} uses F, whereas centred M_{int} uses HF, with $H = I - \mathbf{1}\mathbf{1}^\top/n$.Centring removes each feature's mean across examples. Raw overlap includes contributions from these means and is not a covariance measure. Each calculation retains only its nonzero feature columns.

3.2 Representation trajectories and two geometry observations

To study changes over time, fix a horizon $T > 0$, an ordered set of n probe examples, and K features. Each $F(t) \in \mathbb{R}^{n\times K}$ records their activations at time t, with row i corresponding to the same probe example throughout. We consider the trajectory space

$$\mathcal{X} \;=\; \Big\{\, F : [0,T] \to \mathbb{R}^{n\times K} \;:\; F \text{ absolutely continuous},\ \operatorname{rank} F(t) = K \text{ for all } t \,\Big\}. \tag{3}$$

Absolute continuity excludes jumps and permits differentiation almost everywhere; full column rank requires the features to remain linearly independent.

Each trajectory is observed through its Gram geometry. The *feature-side* and *sample-side* Grams are

$$G_f(t) = F(t)^\top F(t) \in \mathbb{R}^{K\times K}, \qquad G_s(t) = F(t)F(t)^\top \in \mathbb{R}^{n\times n}. \tag{4}$$

These record inner products between features and between examples, respectively. We define the full-geometry observation and the feature-Gram observation as

$$\mathcal{G}(F) = \big(G_f(t), G_s(t)\big)_{t\in[0,T]}, \qquad \mathcal{M}(F) = \big(G_f(t)\big)_{t\in[0,T]}.$$

By (1), the complete M_{int} history is determined by $\mathcal{M}(F)$. The sample-side Gram additionally determines the representation subspace in the fixed sample coordinates. Keeping only $\mathcal{M}(F)$ therefore discards information that may matter for target alignment.

3.3 Target-subspace alignment and the persistent transition

To measure target relevance, fix a nonzero target vector $y \in \mathbb{R}^n$, with y_i corresponding to probe example i. All trajectory comparisons below use the same target and the same ordering of probe examples. Let P_F be the orthogonal projection onto the span of the feature columns. Define target-subspace alignment by

$$\Omega_y(F) \;=\; \frac{\big\|P_F\, y\big\|_2^2}{\|y\|_2^2} \;\in\; [0,1]. \tag{5}$$

Since $P_F = G_s G_s^\dagger$, where $\dagger$ denotes the Moore–Penrose pseudoinverse, G_s determines Ω_y for this fixed target. Alignment also determines the smallest squared error achievable by a linear readout on the probe examples:

$$\min_{w\in\mathbb{R}^K} \big\|Fw - y\big\|_2^2 \;=\; \|y\|_2^2\big(1-\Omega_y(F)\big), \tag{6}$$

Thus Ω_y is the fraction of target energy captured by the representation, and higher alignment means lower optimal linear-readout error. To describe a persistent transition, fix an alignment threshold $\alpha \in (0,1)$ and a minimum persistence duration $\Delta \in (0,T)$, and define

$$\begin{aligned} \tau(\alpha,\Delta) &= \inf\big\{\, t\in[0,T-\Delta] :\; \Omega_y\big(F(s)\big) \ge \alpha \text{ for all } s\in[t,T]\,\big\}, \\ g(\alpha,\Delta) &= \mathbf{1}\big\{\tau(\alpha,\Delta) < \infty\big\}, \end{aligned} \tag{7}$$

Here $\inf\varnothing = \infty$. The event $g(\alpha,\Delta)=1$ occurs when alignment reaches the threshold no later than $T-\Delta$ and remains at or above it until T. We call this a target-geometric transition; its connection to test-defined grokking requires the assumptions in Section 3.6

3.4 Observable grokking ambiguity

An observation $O : \mathcal{X} \to \mathcal{Y}_O$ specifies the information available about each trajectory. The event $g(\alpha,\Delta)$ is observable from O if this information determines whether the event occurs: there exists a function φ such that

$$\varphi\big(O(F)\big) = g(\alpha,\Delta)(F) \qquad \text{for every } F\in\mathcal{X}; \tag{8}$$

Equivalently, trajectories with identical observations must have the same event outcome. We apply this definition to $\mathcal{G}$ and $\mathcal{M}$, and record whether ambiguity remains using

$$\mathrm{OGA}(O) \;=\; \mathbf{1}\Big\{\exists\, F,F'\in\mathcal{X} :\; O(F)=O(F'),\; g(\alpha,\Delta)(F)\neq g(\alpha,\Delta)(F')\Big\}. \tag{9}$$

Thus $\mathrm{OGA}(O)=0$ means the observation determines the event, while $\mathrm{OGA}(O)=1$ means that two trajectories can look identical under O but have different outcomes.

3.5 Dynamic unobservability

Theorem 1 (Dynamic unobservability under the feature-overlap projection)**.** *Given fixed $T>0$, $\alpha\in(0,1)$, $\Delta\in(0,T)$, a nonzero target $y\in\mathbb{R}^n$, and $2\le K\le n-1$, there exist two trajectories $F^{(N)}, F^{(G)}\in\mathcal{X}$ with $F^{(N)}(0)=F^{(G)}(0)$ such that*

$$G_f^{(N)}(t) = G_f^{(G)}(t) \quad \textit{for all } t\in[0,T],$$

—so the two runs share identical normalised overlap matrices, identical complete M_{int} histories, and identical outputs of any history-only detector—while

$$g(\alpha, \Delta)\big(F^{(N)}\big) = 0 \qquad \textit{and} \qquad g(\alpha, \Delta)\big(F^{(G)}\big) = 1.$$

Consequently, no function of the complete M_{int} history determines $g(\alpha, \Delta)$ or $\tau(\alpha, \Delta)$ on $\mathcal{X}$: $\mathrm{OGA}(\mathcal{M}) = 1$. *The full-geometry observation, in contrast, determines Ω_y, $\tau(\alpha, \Delta)$, and $g(\alpha, \Delta)$, so* $\mathrm{OGA}(\mathcal{G}) = 0$.

In the construction of trajectories, one representation subspace rotates towards the target while the other remains orthogonal to it. Their feature-side Gram matrices agree throughout, so the change in target alignment is invisible to M_{int}. The proof is given in Appendix A.2.

Proposition 1 (Dynamic observability under shared non-trivially declining M_{int} histories)**.** *For any absolutely continuous $\rho : [0, T] \to (0, 1)$, the pair in Theorem 1 can be chosen with $M_{\mathrm{int}}\big(F^{(N)}(t)\big) = M_{\mathrm{int}}\big(F^{(G)}(t)\big) = \rho(t)$ for all t. Taking ρ strictly decreasing, the two runs share the entire overlap-decline curve observed empirically during training and still differ in $g(\alpha, \Delta)$.*

Corollary 1 (Feature-Gram-only rule cannot identify the grokking transition)**.** *The incapability is not specific to the scalar M_{int}: no rule—target-aware or otherwise—whose input is any functional of the feature-side Gram path $t \mapsto F(t)^\top F(t)$ (entrywise statistics, spectra, derivatives, integrals, extremal times, or detector outputs) can determine $g(\alpha, \Delta)$ on $\mathcal{X}$.*

Scope. Theorem 1 identifies an information limit over the trajectory space $\mathcal{X}$. The constructed trajectories are not shown to arise from a particular training algorithm; implications for behavioural grokking require Section 3.6. The empirical value of M_{int} as a descriptor and intervention target is examined in Section 5.

3.6 Geometry-to-behaviour assumptions

The unconditional theory above concerns the target-geometric event $g(\alpha, \Delta)$. Connecting it to behavioural grokking—a persistent test-accuracy crossing on the same horizon—requires assumptions that we state explicitly and use nowhere silently.

Assumption 1 (Geometry-to-behaviour bridge)**.** Within the model class and training protocol under study, there are thresholds $\alpha \in (0, 1)$ and $\Delta > 0$ such that a run exhibits the behavioural event if and only if its trajectory satisfies $g(\alpha, \Delta) = 1$, and the behavioural onset time is a function of $\tau(\alpha, \Delta)$.

Assumption 2 (Realizability)**.** Some architecture and optimization procedure in the class under study, with some task and target y, induces a pair of runs with the same initial representation and the same feature-side Gram path whose trajectories differ in $g(\alpha, \Delta)$.

Proposition 2 (Conditional failure of behavioural prediction from overlap)**.** *Under Assumptions 1 and 2, no function of the complete M_{int} history determines behavioural grokking on the realised runs. Without either assumption, no statement about test-defined grokking follows from Theorem 1 alone.*

4 Experimental Protocol

Assumption 1 is a strong bridge that can be hardly established in practice; while Assumption 2 is a concrete target for future analysis of specific optimisers. The following empirical sections therefore treat M_{int} as an informative descriptor of feature-overlap dynamics—and every behavioural claim made below is empirical. The results reflect the gap between empirical results and Theorem 1.

Tasks, model, and observables. Modular addition uses all $113^2 = 12{,}769$ ordered pairs and 113 classes. Nonzero modular multiplication uses $(a, b) \in \{1, \ldots, 112\}^2$, 12,544 examples, and 112 balanced output classes. Inputs concatenate two one-hot vectors. The model is a two-linear-layer ReLU MLP with hidden width 512. The feature matrix is the post-ReLU hidden activation evaluated over the complete training split at each logging checkpoint. Unless stated otherwise, training uses AdamW, learning rate 3×10^{-4}, weight decay 0.1, batch size 64, 10,000 epochs, and logging at the untrained epoch 0 and every 10 completed epochs. Activity thresholds are relative 10^{-8} and absolute 10^{-12}.

We define t_{train} by training accuracy ≥ 0.99 and robust t_{grok} as the first of three consecutive logged test accuracies ≥ 0.95. Non-events are right-censored. Retention is the fraction of post-onset checkpoints at or above 0.95; normalised test AUC covers epochs 1,000–10,000. Curves and reported dispersions use mean $\pm$ sample SD; additional event definitions and uncertainty procedures appear in Appendix B.

Controls freeze hidden features, permute training-example targets while preserving the class histogram and evaluating against true labels, or reduce the addition training split to 10%. Multiplication uses a disclosed pilot-selected configuration, with pilot runs excluded from formal statistics (Appendix B).

For minibatch activations H, we train with $L = L_{\text{CE}} + \beta M_{\text{int}}(H)^2$, using $\beta \in \{-0.17, -0.057, 0, 0.057, 0.17\}$ and five seeds paired by split, initialisation, and minibatch order. Positive β penalises overlap; negative β encourages it. Corrected controls match the hidden-parameter pre-Adam gradient norm while preserving the task-gradient direction; construction and numerical checks appear in Appendix B

The history-only detector alarms after two consecutive decreases in logged raw M_{int} (10 epochs per interval). Its parameters are calibrated on addition seeds 0-4 and frozen for all transfer conditions; implementation and local randomness designs appear in Appendix B.

5 Empirical Results

Two-task replication Figure 2 and Table 1 show an early raw-overlap peak followed by delayed generalization in both modular tasks. For addition, training fit occurs at $t_{\text{train}} = 121.3 \pm 3.5$, the raw peak at $t_{\text{peak}} = 40 \pm 0$, and robust onset at $t_{\text{grok}} = 798.0 \pm 26.0$ epochs; multiplication shows the same ordering, with $t_{\text{train}} = 115.0 \pm 5.3$, raw $t_{\text{peak}} = 40 \pm 0$, and robust $t_{\text{grok}} = 789.0 \pm 35.7$ epochs (Table 1). Thus, the raw peak precedes behavioural onset by hundreds of epochs in these MLP runs and should not be identified with the onset itself (Figure 2; Table 1). Final raw overlap remains nonzero, at 0.395 for addition and 0.370 for multiplication (Table 1).

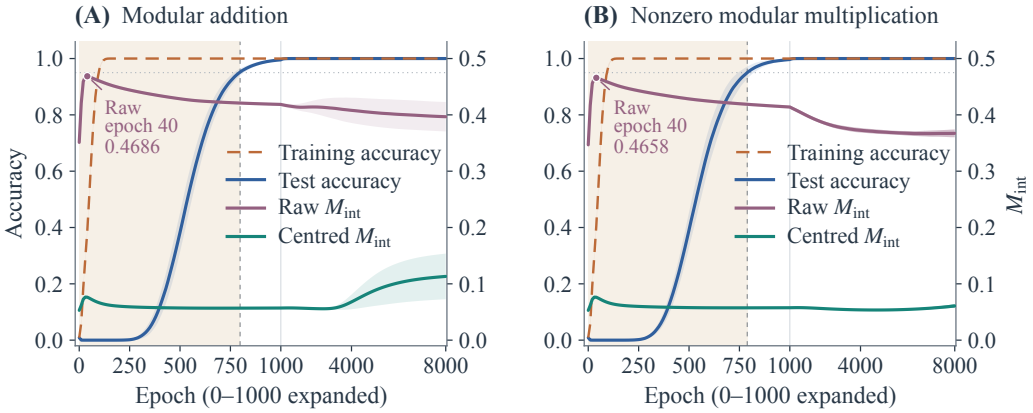

Figure 2: Two-task replication in the tested MLP family. Curves show mean $\pm$ sample SD. Black dashed lines mark mean t_{train}, blue dotted lines mean raw t_{peak}, and green dash–dot lines mean robust t_{grok}. The raw peak precedes the test-defined transition in both tasks; centred overlap is a separate convention.

Table 1: Two-task observational replication in the tested MLP family (protocol A, Figure 2; mean $\pm$ sample SD).

Task	n	t_{train}	raw t_{peak}	robust t_{grok}	final raw / centred
Addition	15	121.3 ± 3.5	40 ± 0	798.0 ± 26.0	0.395 / 0.117
Multiplication	10	115.0 ± 5.3	40 ± 0	789.0 ± 35.7	0.370 / 0.074

Convention sensitivity and non-sufficiency controls Raw and centred M_{int} exhibit different long-horizon behaviour, whereas frozen features produce stationary overlap (Figure 3a). In addition, 12 of 15 centred maxima occur at the 10,000-epoch boundary and are endpoint-censored rather than observed interior peaks (Appendix C.1). Permuted-target and low-data addition runs fit their training targets without crossing the true-label test threshold, despite changes in raw overlap (Figure 3b). Multiplication reproduces this non-sufficiency: none of the ten permuted-target runs robustly groks, and final true-label test accuracy is 0.00899 ± 0.00126, close to chance $1/112$ (Appendix C.3). These controls show that overlap change alone is insufficient for true-label generalization in the tested settings (Figure 3b; Appendix C.3).

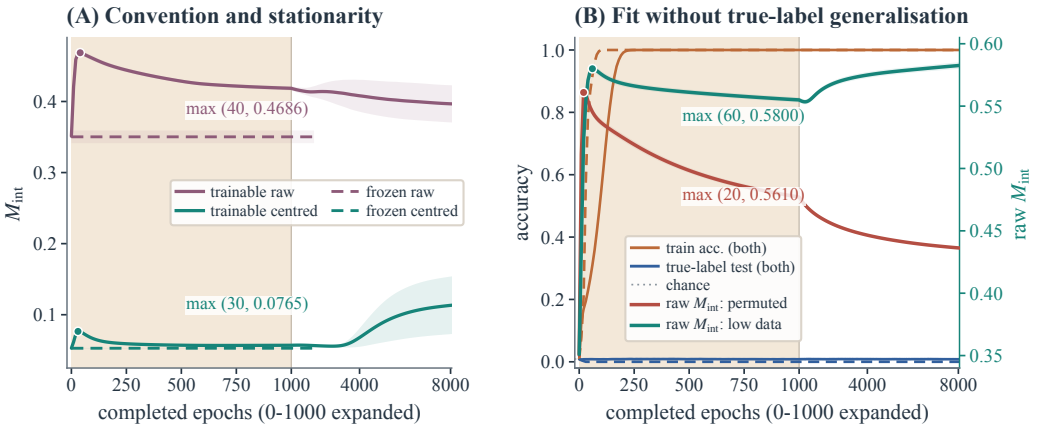

Figure 3: Representation convention and controls. Raw and centred Mint are distinct observables. Frozen features are stationary; target permutation and low data produce overlap changes without true-label threshold crossing.

Mint-penalized training and signed-coefficient controls. Positive coefficients reduce raw overlap, whereas negative coefficients increase it (Figure 4a). Relative to the paired unpenalized baseline (800 ± 29 epochs), $\beta = 0.057$ and 0.17 advance robust onset to 468 ± 13 and 398 ± 13 epochs, respectively (Figure 4b; Table 2). However, every positive-penalty run subsequently falls below the 0.95 threshold: mean retention drops to 0.167 and 0.047, and final accuracy to 0.938 and 0.885, respectively (Table 2). Negative coefficients yield no robust crossing within 10,000 epochs (Figure 4b; Table 2). Corrected gradient-norm-matched controls grok near epoch 856 with near-unit final accuracy and retention, without reproducing the signed overlap endpoints (Figure 4b; Table 2). The results therefore show an onset–retention trade-off and do not support the tested pre-Adam gradient-norm magnitude explanation; gradient direction, Adam moments, and path dependence remain possible (Figure 4; Table 2).

Table 2: Signed overlap interventions and corrected gradient-norm-matched controls (five paired seeds per condition). AUC is normalised over epochs 1,000–10,000. G→U records any post-onset dip below 0.95; duration and endpoint metrics are needed for interpretation.

Condition	Robust grok	Robust t_{grok}	Final test	AUC	Retention	G→U
-0.17	0/5	censored	0.895	0.846	–	0/5
-0.057	0/5	censored	0.920	0.907	–	0/5
0	5/5	800 ± 29	1.000	1.000	1.000	0/5
$+0.057$	5/5	468 ± 13	0.938	0.938	0.167	5/5
$+0.17$	5/5	398 ± 13	0.885	0.892	0.047	5/5
match -0.17	5/5	856 ± 34	1.000	1.000	0.998	4/5
match $+0.17$	5/5	856 ± 40	1.000	1.000	0.999	3/5

Transfer of the fixed overlap-decline detector. The frozen detector alarms at epoch 60 in all ten held-out addition runs and all ten multiplication baselines, with no alarm in frozen-hidden runs (Figure 5a,b; Table 3). However, all addition and multiplication permuted-target runs and all low-data runs also alarm, while none robustly groks by the horizon (Figure 5b; Table 3). Detector time

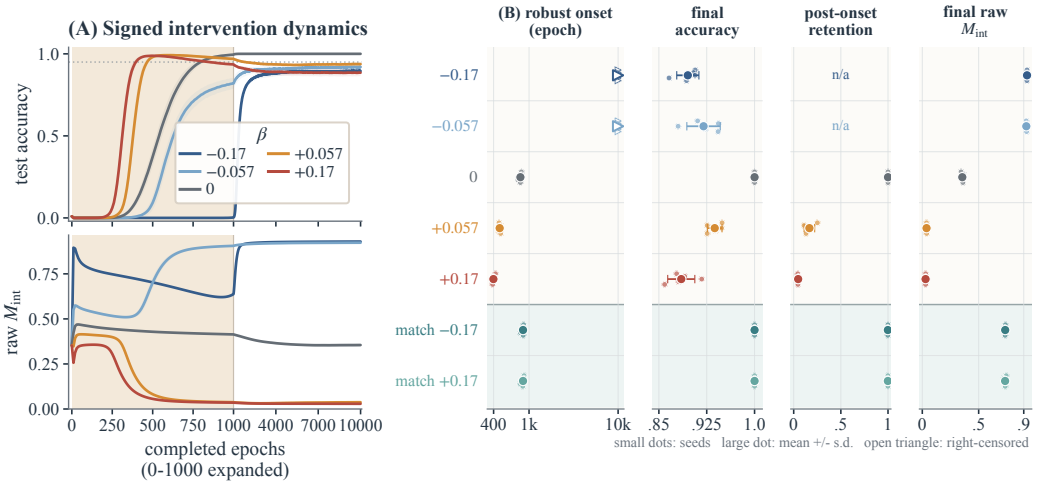

Figure 4: Signed overlap interventions and corrected matched controls. Positive penalties advance robust threshold onset but reduce retention and final accuracy; negative coefficients raise overlap and are right-censored at 10,000 epochs. Open arrows indicate censoring. Gradient-norm-matched controls do not reproduce the signed endpoints.

also responds differently from grokking time: lowering the learning rate to 10^{-4} gives mean detector and robust onset times of 148 and 2438 epochs, respectively; reducing weight decay to 0.03 gives 54 and 2200 epochs (Figure 5a; Table 3). These results support transfer as a detector of early raw-overlap decline, but neither establish alarm sufficiency for grokking nor support identifying detector time with behavioural onset (Figure 5; Table 3).

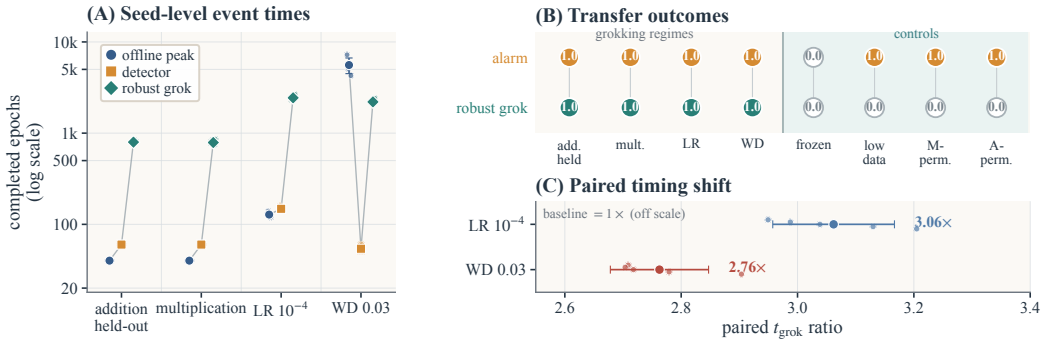

Figure 5: Frozen-detector transfer. Alarms identify an early raw-overlap decline but occur in non-generalising controls and do not track the large condition- or seed-level changes in robust grokking time.

Table 3: Frozen-detector transfer. Every row uses the same frozen detector. Alarms in permuted-target and low-data conditions show that the alarm is not sufficient for generalization.

Condition	Alarm	Mean t_{detect}	Robust grok	Mean robust t_{grok}
Addition held-out	10/10	60	10/10	797
Frozen hidden	0/10	–	0/10	–
Addition permuted	10/10	40	0/10	–
Low data	10/10	80	0/10	–
Multiplication	10/10	60	10/10	789
Multiplication permuted	10/10	40	0/10	–
LR 10^{-4}	5/5	148	5/5	2438
WD 0.03	5/5	54	5/5	2200

6 DISCUSSION AND LIMITATIONS

Three findings are concluded. First, both the theoretical study and empirical results consistently show that feature overlap is not equivalent to the target relevance. Theorem 1 shows constructively that a complete M_{int} history cannot determine the target-geometric transition.Empirically, the permuted-target and low-data controls exhibit changes in representation overlap without true-label generalization, illustrating that overlap change alone is insufficient for generalization in the tested settings (Figure 3b; Appendix C.3).

Second, directional intervention is not unique causation. The signed-β study provides causal evidence that overlap-directed gradients alter onset and retention in the tested system, and matched controls rule out the direct gradient-norm magnitude explanation tested here.(Figure 4; Table 2) They do not eliminate gradient-direction, Adam-moment, or path-dependence alternatives.

Third, representation change is not behavioural prediction. The frozen rule consistently detects early raw-overlap decline in held-out addition runs and multiplication baselines (Figure 5; Table 3). However, it also alarms in non-generalising controls, and its alarm times do not track the large shifts in grokking onset caused by changes in learning rate or weight decay (Figure 5; Table 3). Its alarms therefore indicate overlap decline but do not reliably determine whether or when grokking occurs.

The above findings are limited to two modular tasks, one shallow MLP family, five seeds per intervention, 10-epoch logging,pilot-selected multiplication settings, one detector calibration set, and one local randomness anchor. Protocol B covers one Transformer and one task, uses a separate environment, and is not pooled with the 142 MLP runs. Incomplete protocol-A runtime metadata are disclosed in Appendix E. An independent experiment shows that a one-layer Transformer probe further limits temporal generality: all ten baselines grok, yet the primary-site that raw maximum M_{int} occurs at epoch 0 for all 20 baseline in permuted-target runs. The decline of feature-overlap also persists when generalization occurs given intentionally wrongly labeled ground truth during training. Thus the MLP interior early peak does not transfer to this probe (Appendix D).

7 CONCLUSION

We formalised whether representation geometry provides a reliable observable of grokking as a dynamic observability question. Our dynamic unobservability theorem shows that identical complete M_{int} histories can accompany different target-geometric outcomes; extending this result to test-defined grokking requires explicit geometry-to-behaviour assumptions. Nevertheless, M_{int} provides a label-free descriptor of representation-overlap dynamics. Across two modular tasks in the tested MLP family, it captures reproducible early overlap changes, while the frozen detector identifies overlap decline without reliably determining whether or when grokking occurs. Moreover, M_{int}-penalized training within the tested positive-coefficient range advances generalization onset, but reduces post-onset retention and final accuracy.Together, these results clarify the limits of feature overlap as an observable of grokking and identify representation overlap as a practical intervention target, while showing that earlier generalization onset need not lead to sustained generalization.

Reproducibility statement. All reported values are rebuilt from machine-readable histories by a frozen evidence script. The supplement provides complete definitions, proofs, configurations, event and censoring rules, per-run results, paired-hash checks, detector provenance, numerical integrity checks, figure-generation scripts, and known artifact gaps. No numerical value is inferred from an image.

AI Use Statement. Generative AI tools were used to assist with manuscript restructuring, language editing, literature discovery, analysis-script drafting, consistency checks between reported values and machine-readable artifacts, figure planning and generation, and interpretation review. Generative AI tools did not execute the original training runs. The authors take responsibility for the accuracy and integrity of the final manuscript and all accompanying artifacts, including AI-assisted text, code, and figures.

REFERENCES

Ryan Prescott Adams and David J. C. MacKay. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*, 2007.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022.

Bilal Chughtai, Lawrence Chan, and Neel Nanda. A toy model of universality: Reverse engineering how networks learn group operations. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 6243–6267, 2023.

Michael Cogswell, Faruk Ahmed, Ross Girshick, Larry Zitnick, and Dhruv Batra. Reducing overfitting in deep networks by decorrelating representations. In *International Conference on Learning Representations*, 2016.

Shreel Golwala. ILDR: Geometric early detection of grokking. *arXiv preprint arXiv:2604.20923*, 2026.

Jianliang He, Leda Wang, Siyu Chen, and Zhuoran Yang. On the mechanism and dynamics of modular addition: Fourier features, lottery ticket, and grokking. *arXiv preprint arXiv:2602.16849*, 2026.

Gunner Levi Howe. Structure-specific representational priors causally control the grokking delay. *arXiv preprint arXiv:2607.04333*, 2026.

Maksim A. Kazanskii. How to tame grokking: Representation geometry as a control signal. *arXiv preprint arXiv:2607.11666*, 2026.

Truong Xuan Khanh. At-grok is not converged: A measurement-validity audit for grokking representation metrics. *arXiv preprint arXiv:2607.06639*, 2026.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529, 2019.

Tanishq Kumar, Blake Bordelon, Samuel J. Gershman, and Cengiz Pehlevan. Grokking as the transition from lazy to rich training dynamics. In *The Twelfth International Conference on Learning Representations*, 2024.

Jaerin Lee, Bong Gyun Kang, Kihoon Kim, and Kyoung Mu Lee. Grokfast: Accelerated grokking by amplifying slow gradients. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Noam Itzhak Levi, Alon Beck, and Yohai Bar-Sinai. Grokking in linear estimators: A solvable model that groks without understanding. In *International Conference on Learning Representations*, 2024.

Ziming Liu, Ouail Kitouni, Niklas S. Nolte, Eric Michaud, Max Tegmark, and Mike Williams. Towards understanding grokking: An effective theory of representation learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Kaifeng Lyu, Jikai Jin, Zhiyuan Li, Simon Shaolei Du, Jason D. Lee, and Wei Hu. Dichotomy of early and late phase implicit biases can provably induce grokking. In *International Conference on Learning Representations*, 2024.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations*, 2023.

E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954.

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.

Lucas Prieto, Melih Barsbey, Pedro A. M. Mediano, and Tolga Birdal. Grokking at the edge of numerical stability. In *The Thirteenth International Conference on Learning Representations*, 2025.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, 2017.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *15th European Signal Processing Conference*, pp. 606–610, 2007.

Achyuthan Sivasankar. Circuit synchronization precedes generalization: A causal precursor to grokking. *arXiv preprint arXiv:2606.12966*, 2026.

Vimal Thilak, Etai Littwin, Shuangfei Zhai, Omid Saremi, Roni Paiss, and Joshua Susskind. The slingshot effect: A late-stage optimization anomaly in adaptive gradient methods. *Transactions on Machine Learning Research*, 2024.

Yuandong Tian. Provable scaling laws of feature emergence from learning dynamics of grokking. In *The Fourteenth International Conference on Learning Representations*, 2026.

Xuan Khanh Truong, Quynh Hoa Truong, Duc Trung Luu, and Thanh Duc Phan. Spectral entropy collapse as an empirical signature of delayed generalisation in grokking. *arXiv preprint arXiv:2604.13123*, 2026.

Vikrant Varma, Rohin Shah, Zachary Kenton, János Krámár, and Ramana Kumar. Explaining grokking through circuit efficiency. *arXiv preprint arXiv:2309.02390*, 2023.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12310–12320, 2021.

A Proofs and Formal Statements

Recent two-layer analysis models late-stage hidden-unit interaction through inverse regularised Gram geometry (Tian, 2026). This motivates monitoring pairwise representation overlap, but it does not identify M_{int} with that model's exact update or energy. To state only the algebraic quantity controlled by our observable, define the transparent pairwise surrogate

$$E_{\mathrm{pair}}(F) = -\frac{\lambda}{2}\sum_{j\neq k}(f_j^\top f_k)^2 = -\frac{\lambda}{2}\sum_{j\neq k} G_{jj}G_{kk}C_{jk}^2, \tag{10}$$

with positive surrogate scale $\lambda > 0$ (distinct from the alignment threshold α). If $0 < g_{\min} \leq G_{jj} \leq g_{\max}$, then

$$\frac{\lambda}{2} g_{\min}^2\, K(K-1)\, M_{\mathrm{int}}^2 \;\leq\; |E_{\mathrm{pair}}| \;\leq\; \frac{\lambda}{2} g_{\max}^2\, K(K-1)\, M_{\mathrm{int}}^2. \tag{11}$$

Equation (11) is exact for this surrogate but can be loose under feature-norm imbalance. It is not a claim that the actual AdamW trajectory follows the surrogate flow. Complete statements, a local directional-dominance result, an exact descent result for a normalised angular surrogate, and proofs appear in Appendix A. Throughout this appendix, define $E_{\mathrm{pair}}(F) := -\frac{\lambda}{2}\sum_{j\neq k}(f_j^\top f_k)^2$ and $R_{\mathrm{int}}(F) := M_{\mathrm{int}}(F)^2$, with penalty coefficient $\lambda > 0$ distinct from the alignment threshold α of Section 3. The appendix is organised as follows: A.1 collects exact algebraic statements about the observable; A.2 proves the dynamic unobservability theorem, the shared-decline construction, and their consequences; A.3 proves the conditional geometry-to-behaviour proposition; A.4 states the static non-identifiability precursor; A.5 collects surrogate-flow propositions and the finite-sample detector guarantee. The flow propositions concern deliberately simplified surrogate dynamics and are not identities for AdamW or for the exact inverse-Gram interaction in Tian (2026).

Table 4: Structural properties and empirical interpretation. Exact statements apply to retained nonzero feature columns.

Property	Statement	Interpretation
Range	$0 \le M_{\mathrm{int}} \le 1$	Common scale across checkpoints.
Permutation invariance	$M_{\mathrm{int}}(F\Pi) = M_{\mathrm{int}}(F)$	Independent of hidden-unit ordering.
Feature-wise scale invariance	$M_{\mathrm{int}}(FD) = M_{\mathrm{int}}(F)$ for nonsingular diagonal D	Retains angular rather than magnitude information.
Orthogonality	$M_{\mathrm{int}} = 0$ iff columns are pairwise orthogonal	Zero is exact but need not be attainable.
PR identity	$r_{\mathrm{PR}} = K/[1 + (K-1)M_{\mathrm{int}}^2]$	Effective rank is an equivalent reparameterisation.
Rank floor	$M_{\mathrm{int}}^2 \ge (K-r)/[r(K-1)]$ when $r < K$	Finite rank can force nonzero residual overlap.
Dynamic unobservability	Identical complete M_{int} histories can accompany opposite $g(\alpha, \Delta)$	The overlap history does not determine the transition (Theorem 1).

A.1 Algebraic structure of the observable

Proposition 3 (Quadratic selection for the pairwise surrogate)**.** *For $K \ge 2$, consider $O_{\phi,K}(F)^2 = \frac{1}{K(K-1)} \sum_{j \ne k} \phi(C_{jk})$, where $\phi : [-1, 1] \to [0, \infty)$ is even, $\phi(0) = 0$, and $\phi(1) = 1$. Suppose that whenever all feature norms are equal, $|E_{\mathrm{pair}}(F)|$ is proportional to $O_{\phi,K}(F)^2$ for every feasible feature matrix, with a proportionality factor independent of pairwise cosines. Then $\phi(c) = c^2$ and $O_{\phi,K} = M_{\mathrm{int}}$.*

Proof. Apply the assumption at $K = 2$. Equal-norm pairs can realise every $c \in [-1, 1]$. Their pairwise-surrogate magnitude is proportional to c^2, while $O_{\phi,2}^2 = \phi(c)$. Proportionality for all c, together with $\phi(1) = 1$, gives $\phi(c) = c^2$.

Proposition 4 (Basic properties)**.** *For every F with nonzero columns: (i) $0 \le M_{\mathrm{int}}(F) \le 1$; (ii) $M_{\mathrm{int}}(F\Pi) = M_{\mathrm{int}}(F)$ for every permutation matrix Π; (iii) $M_{\mathrm{int}}(FD) = M_{\mathrm{int}}(F)$ for every diagonal D with nonzero entries; and (iv) $M_{\mathrm{int}}(F) = 0$ if and only if the columns of F are pairwise orthogonal.*

Proof. Cauchy–Schwarz gives $|C_{jk}| \le 1$, establishing the range. A permutation only reorders off-diagonal entries. Under $F \mapsto FD$, each cosine changes by $\mathrm{sign}(d_j d_k)$, which vanishes after squaring. Finally, a sum of nonnegative squared cosines is zero exactly when each off-diagonal inner product is zero.

Proposition 5 (Participation-ratio identity and rank floor)**.** *Let $U = [u_1, \ldots, u_K]$ contain unit feature directions, let $r = \mathrm{rank}(U)$, and define $r_{\mathrm{PR}}(U) = \frac{\mathrm{tr}(UU^\top)^2}{\|UU^\top\|_F^2}$. Then equation* (2) *holds and $M_{\mathrm{int}}(F)^2 \ge \max\left\{0, \frac{K-r}{r(K-1)}\right\}$.*

Proof. Every column of U has unit norm, so $\mathrm{tr}(UU^\top) = K$. Moreover, $\|UU^\top\|_F^2 = \|U^\top U\|_F^2 = K + K(K-1)M_{\mathrm{int}}^2$, which gives the identity. Let $\lambda_1, \ldots, \lambda_r$ be the nonzero eigenvalues of $UU^\top$. Since $\sum_i \lambda_i = K$, Cauchy–Schwarz yields $\sum_i \lambda_i^2 \ge K^2/r$. Substitution gives the floor.

Proposition 6 (Norm-conditioned pairwise-surrogate bounds)**.** *If $0 < g_{\min} \le G_{jj} \le g_{\max} < \infty$ for every j, then equation* (11) *holds.*

Proof. For each ordered pair $j \ne k$, $g_{\min}^2 \le G_{jj}G_{kk} \le g_{\max}^2$. Multiply by $C_{jk}^2 \ge 0$, sum, and use $\sum_{j \ne k} C_{jk}^2 = K(K-1)M_{\mathrm{int}}^2$.

A.2 DYNAMIC UNOBSERVABILITY: PROOFS AND CONSEQUENCES

Proof of Theorem 1. The construction proceeds in six numbered steps.

1. *Target direction and orthonormal basis.* Set $\hat{y} = y/\|y\|_2$. Since $K \le n-1$, the orthogonal complement $\hat{y}^\perp$ contains K orthonormal vectors $q_1, \dots, q_K$. Let $Q_N = [q_1, \dots, q_K]$, so $Q_N^\top Q_N = I_K$ and $Q_N^\top \hat{y} = 0$.

2. *Delayed rotation.* Because $T - \Delta > 0$, choose $0 < t_d < t_* \le T - \Delta$ and an absolutely continuous, nondecreasing $\theta : [0, T] \to [0, \pi/2]$ with $\theta \equiv 0$ on $[0, t_d]$, $\theta(t_*) = \pi/2$, and $\theta \equiv \pi/2$ on $[t_*, T]$. Define

$$q_1^{(G)}(t) = \cos\theta(t)\, q_1 + \sin\theta(t)\, \hat{y}, \qquad Q_G(t) = \big[q_1^{(G)}(t),\, q_2, \dots, q_K\big].$$

   Since $q_1^{(G)}(t)$ is a unit vector in $\operatorname{span}\{q_1, \hat{y}\}$ and is orthogonal to $q_2, \dots, q_K$ (which lie in $\hat{y}^\perp$), every $Q_G(t)$ has orthonormal columns; and $\theta(0) = 0$ gives $Q_G(0) = Q_N$.

3. *Shared feature-side geometry.* Let $S : [0, T] \to \mathrm{SPD}(K)$ be any absolutely continuous path in the symmetric positive-definite cone (Proposition 1 instantiates it) and set

$$F^{(N)}(t) = Q_N\, S(t), \qquad F^{(G)}(t) = Q_G(t)\, S(t).$$

   Each $S(t)$ is invertible, so both trajectories retain full column rank at every t; $Q_G(0) = Q_N$ gives the common initial representation, so $F^{(N)}, F^{(G)} \in \mathcal{X}$.

4. *Identical complete M_{int} histories.* Using $Q_N^\top Q_N = Q_G^\top Q_G = I_K$ and $S = S^\top$,

$$F^{(N)}(t)^\top F^{(N)}(t) = S(t)^\top S(t) = S(t)^2 = S(t)^\top Q_G(t)^\top Q_G(t) S(t) = F^{(G)}(t)^\top F^{(G)}(t)$$

   for every t. The feature-side Grams coincide, hence so do the normalised overlap matrices $C(t)$, the values $M_{\mathrm{int}}(t)$, and every functional of the complete history: derivatives, integrals, peaks, decline times, and all history-only detector outputs.

5. *Opposite target-relevant outcomes.* Since $S(t)$ is invertible, $\operatorname{col} F = \operatorname{col} Q$ and $P_F = QQ^\top$. For the first trajectory the column space lies in $\hat{y}^\perp$, so $\Omega_y(F^{(N)}(t)) = 0$ for all t and $g(\alpha, \Delta)(F^{(N)}) = 0$. For the second, only $q_1^{(G)}(t)$ has a component along $\hat{y}$, of size $\sin\theta(t)$, so

$$\Omega_y\big(F^{(G)}(t)\big) = \sin^2\theta(t).$$

   Early in training $\theta = 0$ and the run carries no target alignment; after the delay, θ increases to $\pi/2$ and stays there. Once $\theta(t) \ge \arcsin\sqrt{\alpha}$—which first happens no later than $t_* \le T - \Delta$—alignment is at least α and remains so through T; hence $\tau(\alpha, \Delta) \le T - \Delta$ and $g(\alpha, \Delta)(F^{(G)}) = 1$.

6. *Contradiction and determinacy.* Suppose some φ satisfied $\varphi(\mathcal{M}(F)) = g(\alpha, \Delta)(F)$ for all $F \in \mathcal{X}$. By Step 4, $\mathcal{M}(F^{(N)}) = \mathcal{M}(F^{(G)})$, so φ must return the same value on both, yet Step 5 gives $g(\alpha, \Delta)(F^{(N)}) \ne g(\alpha, \Delta)(F^{(G)})$—a contradiction. Hence no such φ exists and $\mathrm{OGA}(\mathcal{M}) = 1$. Conversely, $\mathcal{G}$ contains G_s, and $P_F = G_s G_s^\dagger$ determines Ω_y at every t, hence $\tau(\alpha, \Delta)$ and $g(\alpha, \Delta)$; therefore g is constant on every $\mathcal{G}$-fiber and $\mathrm{OGA}(\mathcal{G}) = 0$.

Proof of Proposition 1. In two numbered steps.

1. *The shared SPD path.* Let $R(t) = (1 - \rho(t))I_K + \rho(t)\mathbf{1}\mathbf{1}^\top$, whose diagonal entries equal 1 and whose off-diagonal entries equal $\rho(t)$. Its eigenvalues are $1 - \rho(t)$ with multiplicity $K - 1$ and $1 + (K-1)\rho(t)$, both strictly positive for $\rho(t) \in (0, 1)$; hence $R(t) \in \mathrm{SPD}(K)$, $S(t) = R(t)^{1/2}$ is an absolutely continuous SPD path, and we instantiate Step 3 of Theorem 1 with this S.

2. *Shared declining histories.* Both trajectories then have $G_f(t) = S(t)^2 = R(t)$, whose normalisation has unit diagonal and off-diagonal entries $\rho(t)$, so $M_{\mathrm{int}}(t) = \rho(t)$ for both runs at every t. Choosing ρ strictly decreasing gives the shared non-trivial decline, while Steps 5 and 6 of the theorem are unchanged.

Proof of Corollary 1. Any functional of the feature-side Gram path takes equal values on $F^{(N)}$ and $F^{(G)}$ by Step 4 of Theorem 1, while $g(\alpha, \Delta)$ separates them by Step 5; the argument of Step 6 applies verbatim.

A.3 Geometry-to-behaviour assumptions: proof

Proof of Proposition 2. Under Assumption 2 there is a realised pair of runs with a common initial representation, a common feature-side Gram path, and differing $g(\alpha, \Delta)$; by the computation of Step 4 in Theorem 1, their complete M_{int} histories coincide. Under Assumption 1, differing $g(\alpha, \Delta)$ entails differing behavioural outcomes. Any function of the M_{int} history would therefore have to return two different behavioural verdicts on one input, which is impossible. Conversely, without Assumption 1, $g(\alpha, \Delta)$ and the test-accuracy event need not coincide; and without Assumption 2, the constructed trajectories need not be induced by any training procedure in the class. In either case no unconditional behavioural statement follows from Theorem 1 alone.

A.4 Static non-identifiability (precursor result)

Proposition 7 (Feature overlap does not identify target relevance)**.** *Let $y \in \mathbb{R}^n$ be nonzero and $2 \le K \le n-1$. Define $A(F, y) = \frac{1}{K}\sum_{j=1}^{K} \frac{|y^\top f_j|^2}{\|y\|_2^2 \|f_j\|_2^2}$. There exist $F_1, F_2 \in \mathbb{R}^{n\times K}$ such that $M_{\mathrm{int}}(F_1) = M_{\mathrm{int}}(F_2) = 0$, but $A(F_1, y) = 0$ and $A(F_2, y) = 1/K$.*

Proof. Set $e_1 = y/\|y\|_2$ and choose orthonormal $e_2, \dots, e_{K+1} \in y^\perp$. Let $F_1 = [e_2, \dots, e_{K+1}]$ and $F_2 = [e_1, e_2, \dots, e_K]$. Both have orthonormal columns and hence zero M_{int}. Every column of F_1 is orthogonal to y, whereas exactly one column of F_2 is aligned with y.

A.5 Surrogate flows and detector guarantees

Proposition 8 (Local descent under directional dominance for a surrogate flow)**.** *Suppose on an interval I that a differentiable path satisfies $\dot{W} = v_{\mathrm{int}} + v_{\mathrm{rem}}$ and $\langle \nabla_W R_{\mathrm{int}}, v_{\mathrm{int}} \rangle \le -a(t)$, $|\langle \nabla_W R_{\mathrm{int}}, v_{\mathrm{rem}} \rangle| \le b(t)$, with $a(t) > b(t) \ge 0$ almost everywhere. Then $\frac{d}{dt} R_{\mathrm{int}}(W(t)) \le -[a(t) - b(t)] < 0$ almost everywhere. If $M_{\mathrm{int}} > 0$, then M_{int} is strictly decreasing on the interval.*

Proof. Apply the chain rule to $\dot{W} = v_{\mathrm{int}} + v_{\mathrm{rem}}$ and add the two directional contributions.

Proposition 9 (Exact descent for a normalised angular surrogate flow)**.** *Let $U = (u_1, \dots, u_K)$ lie on a product of unit spheres and define*

$$E_{\mathrm{ang}}(U) = -\frac{\lambda}{2} \sum_{j \ne k} (u_j^\top u_k)^2 = -\frac{\lambda}{2} K(K-1) R_{\mathrm{int}}(U).$$

Under Riemannian gradient ascent $\dot{U} = \operatorname{grad} E_{\mathrm{ang}}(U)$, $\frac{d}{dt} R_{\mathrm{int}}(U(t)) = -\frac{\lambda}{2} K(K-1) \| \operatorname{grad} R_{\mathrm{int}}(U(t))\|^2 \le 0$. Equality holds at stationary points, which need not have zero M_{int}.

Proof. Because $\operatorname{grad} E_{\mathrm{ang}} = -(\lambda/2) K(K-1) \operatorname{grad} R_{\mathrm{int}}$, take the Riemannian inner product with $\operatorname{grad} R_{\mathrm{int}}$.

For a logged sequence Z_t, define $\hat{s}_t(L) = [Z_t - Z_{t-L}]/L$ and

$$\hat{\nu} = \inf\{t \ge L + q - 1 : \hat{s}_{t-r}(L) < -\tau \text{ for every } r = 0, \dots, q-1\}. \tag{12}$$

Theorem 2 (Finite-sample detector guarantee)**.** *Assume $Z_t = m_t + \epsilon_t$ for $t = 0, \dots, T$, where ϵ_t are independent, mean-zero, σ-sub-Gaussian variables. Let $s_t(L) = [m_t - m_{t-L}]/L$. Suppose that for a change time ν, margin $\gamma > 0$, and threshold $\tau \geq 0$, with $\nu + L + q - 1 \leq T$, $s_t(L) \geq -\tau + \gamma$ $(L \leq t < \nu)$, and $s_t(L) \leq -\tau - \gamma$ $(\nu + L \leq t \leq \nu + L + q - 1)$. Let $p_\gamma = \exp[-L^2\gamma^2/(4\sigma^2)]$. Then the detector in equation* (12) *satisfies $P(\hat{\nu} < \nu) \leq (\nu - L)^+ p_\gamma$ and $P(\hat{\nu} > \nu + L + q - 1) \leq q\, p_\gamma$. Consequently, with probability at least $1 - [(\nu - L)^+ + q]p_\gamma$, it raises no pre-change alarm and detects the new regime with delay at most $L + q - 1$ logging intervals.*

Proof. For every $t \geq L$, $\hat{s}_t(L) - s_t(L) = \frac{\epsilon_t - \epsilon_{t-L}}{L}$. The numerator is $\sqrt{2}\sigma$-sub-Gaussian, so a one-sided deviation of magnitude γ has probability at most p_γ. A union bound over pre-change slope tests gives the first result. Once the slope window lies entirely after the change, failure of any of the first q required tests has probability at most p_γ; another union bound gives the delay result. Independence between overlapping slope tests is not required for either union bound.

B Complete Experimental Methods

Architectures. Both tasks use a two-linear-layer MLP: an input-to-hidden affine map, ReLU, and a hidden-to-logit affine map. The feature matrix is the post-ReLU hidden activation over the complete training split. Hidden width is 512. Addition uses input dimension 226 and 113 outputs; nonzero multiplication uses input dimension 224 and 112 outputs.

Datasets and splits. Addition enumerates all ordered pairs $(a, b) \in \{0, \dots, 112\}^2$ with label $(a + b) \bmod 113$. Multiplication enumerates $(a, b) \in \{1, \dots, 112\}^2$ with label $((ab) \bmod 113) - 1$; every class occurs 112 times. Formal baselines use 50% training data. The low-data control uses 10%. Split indices are saved per run. In the target-permutation controls, a random permutation of the training examples is applied to the target vector; this preserves the empirical class histogram and is not a remapping of class identities.

Optimisation. The common configuration is AdamW, learning rate 3×10^{-4}, weight decay 0.1, batch size 64, evaluation batch size 1024, 10,000 completed epochs, and no early stopping. P0-3 changes one hyperparameter at a time: learning rate 10^{-4} or weight decay 0.03. The formal configuration and the untrained epoch-0 state are stored in each run directory. The first-stage baseline was executed in Python 3.9.18/PyTorch 1.12.1 on CPU; P0-4 records a complete RTX 4090 environment snapshot. Cross-environment means are not interpreted as treatment effects; all causal and paired comparisons use within-protocol matched environments.

Observable computation. At each logging point, column norms are computed in double precision. A column is active when its norm exceeds the maximum of a relative threshold 10^{-8} times the largest norm and an absolute threshold 10^{-12}. Raw and centred observables, direct and formula-based PR ranks, norm-conditioned pairwise-surrogate quantities, active counts, and a legacy target-alignment statistic are logged. Across all 142 protocol-A formal runs, the maximum PR-identity discrepancy is 1.71×10^{-13} and the minimum active count is 361; all logged pairwise-surrogate-bound Boolean checks pass.

Event definitions. First crossing uses a single logged point. Robust crossing requires three consecutive points at or above 0.95. Stable entry is the earliest logged point after which all remaining test accuracies are at or above 0.95. Last exit is the final subsequent point below 0.95 after the first crossing. Post-onset retention is the fraction of logged points at or above 0.95 from robust onset to the horizon. Normalised AUC uses trapezoidal integration over epochs 1,000–10,000 divided by 9,000. Right-censored event times remain missing; they are never replaced by the horizon.

P0-1 intervention and matching. For minibatch activations $H \in \mathbb{R}^{B \times K_a}$ after active-feature filtering, columns are normalised and $R_B = \frac{\|U^\top U\|_F^2 - K_a}{K_a(K_a - 1)} = M_{\text{int}}(H)^2$. The objective is $L_{\text{CE}} + \beta R_B$. Pilot seeds not used in formal analysis selected $\beta_1 = 0.057$ and $\beta_2 = 0.17$ by prespecified early gradient ratios. Every signed condition uses seeds 0–4 with identical split, initial parameters, and first 100 minibatches. For the corrected v1.1 matched controls, if g_T is the hidden-parameter task

gradient and g_R the regulariser gradient, the control replaces g_T by $g_{\text{match}} = \frac{\|g_T+\beta g_R\|_2}{\|g_T\|_2+\epsilon}\, g_T$ before AdamW. Output-layer task gradients remain unscaled. Mean and maximum minibatch relative errors are logged; all corrected formal points are below 10^{-6}. Old pre-v1.1 matched runs are retained only as superseded artifacts and are excluded from every paper statistic.

P0-2 configuration selection. Six candidate configurations were ordered in advance. A phase-1 pilot used seed 9200 for 1,500 epochs; candidates that fitted the training set continued with seeds 9200 and 9201 to 10,000 epochs. The first configuration in the fixed order with at least one robust grok event and a train-to-grok delay of at least 200 epochs was selected. C0 (the addition hyperparameters) met the rule. Pilot runs are disclosed but excluded from formal 10-seed statistics.

Detector calibration and transfer. For a logged sequence Z_t, $\hat{s}_t(L) = (Z_t - Z_{t-L})/L$. The alarm is the first time q consecutive slopes are below $-\tau$. Calibration on addition seeds 0–4 selected $L = 1$ logging interval, $\tau = 0$, and $q = 2$; one interval equals 10 completed epochs. The detector configuration was then frozen; no target-condition calibration, rescaling, interpolation, smoothing, or early-point deletion was performed. "Alarm without grokking" in target-permutation or low-data conditions is not called a false alarm, because the detector target is raw-overlap decline. Frozen hidden is the dedicated no-change false-alarm control.

P0-4 design. The one-factor-at-a-time arms fix two random sources and vary the third: model initialisation, data split, or minibatch order. Each arm has eight runs, plus eight fresh coupled-reference runs in the same RTX 4090 environment. Hash checks verify that fixed sources remain byte-identical and the intended source changes. The frozen seed matrix is retained and accompanied by an execution-environment addendum.

Protocol B: preregistered Transformer probe—model and task. The probe uses a one-layer Transformer with explicit Q/K/V causal attention (four heads, $d_{\text{head}} = 32$), $d_{\text{model}} = 128$, $d_{\text{mlp}} = 512$, ReLU activations, no LayerNorm, and no dropout (227,313 parameters; asserted by the unit test `test_protocol.py`). Inputs are length-3 token sequences $[a, b, =]$ over vocabulary $\{0, \ldots, 112\}$ plus the equals token; logits are read from the final token. The task is modular addition mod 113 (all $113^2 = 12{,}769$ ordered pairs) with a 50% train split. These and all protocol-B settings are fixed in the frozen configuration file `results/protocol/formal_config.json`.

Protocol B: training, logging, and observables. Training uses AdamW, learning rate 3×10^{-4}, weight decay 0.1, betas $(0.9, 0.999)$, batch size 64, a horizon of 10,000 completed epochs, and logging at the untrained epoch 0 and every 10 completed epochs. Training runs in FP32; the observable is computed in float64. At each logging point, raw and centred M_{int} are evaluated at three preregistered sites of the final token over the complete training split: the 512-dimensional post-ReLU feed-forward activation (`mlp_post`; the primary, MLP-analogue site), the 128-dimensional residual stream (`resid_post`), and the 128-dimensional attention output (`attn_out`). Raw and centred conventions filter active channels independently, with thresholds relative 10^{-8} and absolute 10^{-12}, identical to protocol A. Event thresholds (0.99 train, 0.95 test, three robust points) match the definitions of protocol A; right-censoring is preserved identically.

Protocol B: seeds and paired structure. Formal seeds are 0–9 for both conditions. Seed s fixes the bundle (run s; split $10000+s$; init $20000+s$; loader $30000+s$; target-permutation $40000+s$). For each seed, the `permuted_target` run shares the split, true targets, initial state, and loader order of the `baseline` run; the only preregistered difference is a permutation of the training-example label vector that preserves the empirical class histogram. Pilot seeds are 9300 and 9301 and are excluded from formal statistics.

Protocol B: preregistration and freezing. Configuration selection read only the three behaviour columns (completed epochs, train accuracy, test accuracy) before freezing; the selection record `pilot_selection.json` logs that `selection_reads_M_int` is false. Candidates were ordered T0, T1, T2; T0 passed on behaviour (both pilot seeds reached train accuracy at least 0.99, at least one robust grok, and a train-to-grok delay of at least 200 epochs: robust grok at epochs 590 and 920 with delays 460 and 740), and selection stopped there. The formal launcher verifies the

frozen hash before every run, and the analyser rejects any incomplete run, grid mismatch, or hash mismatch.

Protocol B: environment. The probe was executed in a separate environment from protocol A (Python 3.12.3, PyTorch 2.8.0+cu128, CUDA), recorded in each run's `metadata.json`. In line with the cross-environment rule above, no cross-protocol mean is interpreted as a treatment effect, and no protocol-B statistic is pooled with the 142 protocol-A runs.

C Additional Results

C.1 Representation convention and structural checks

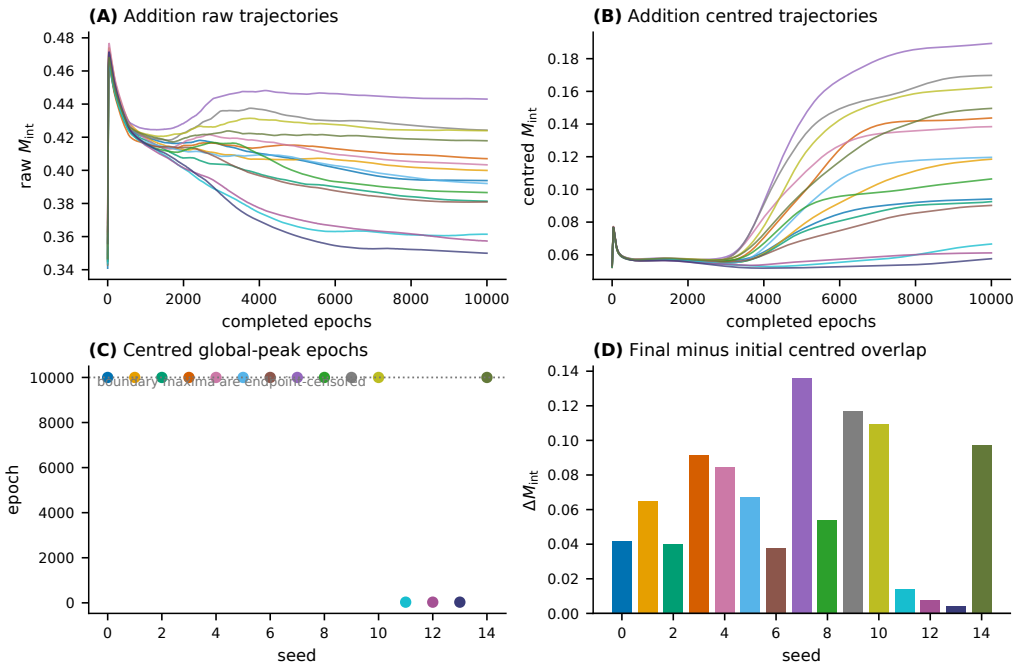

Figure 6: Per-seed addition raw and centred trajectories. Centred maxima at the training boundary are endpoint-censored, not observed interior peaks.

C.2 INTERVENTION DETAILS

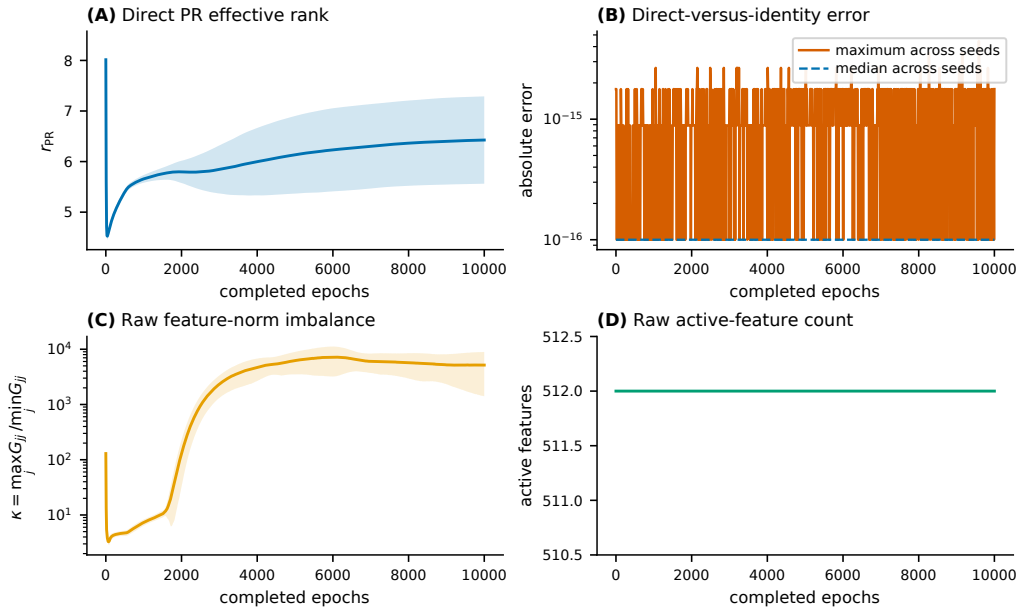

Figure 7: Structural and numerical checks. Direct and formula PR rank agree to numerical precision. Norm imbalance explains why pairwise-surrogate bounds may be loose; all baseline features remain active.

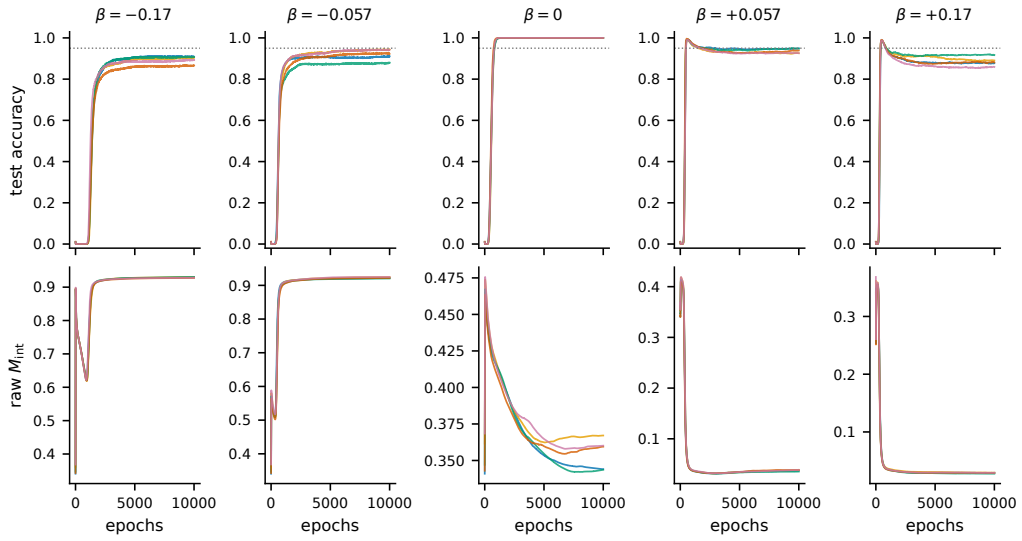

Figure 8: All P0-1 signed-intervention test-accuracy and raw-overlap trajectories. No failed or censored formal seed is removed.

Table 5: P0-1 per-run outcomes.

Condition	Seed	t_{train}	$t_{\text{grok}}^{\text{robust}}$	Final	AUC	Retention	Final raw / centred
beta_neg_strong	0	650	cens.	0.9070	0.856	–	0.9292 / 0.7391
beta_neg_strong	1	660	cens.	0.9040	0.847	–	0.9306 / 0.7426
beta_neg_strong	2	650	cens.	0.9067	0.857	–	0.9305 / 0.7096
beta_neg_strong	3	650	cens.	0.8659	0.816	–	0.9275 / 0.7205
beta_neg_strong	4	650	cens.	0.8933	0.851	–	0.9282 / 0.7336
beta_neg_weak	0	150	cens.	0.9107	0.904	–	0.9218 / 0.7563
beta_neg_weak	1	170	cens.	0.9433	0.928	–	0.9247 / 0.7662
beta_neg_weak	2	160	cens.	0.8802	0.869	–	0.9205 / 0.7354
beta_neg_weak	3	160	cens.	0.9231	0.908	–	0.9217 / 0.7634
beta_neg_weak	4	160	cens.	0.9428	0.928	–	0.9250 / 0.7855
beta_pos_strong	0	90	420	0.8778	0.885	0.043	0.0292 / 0.0190
beta_pos_strong	1	90	400	0.8905	0.904	0.050	0.0295 / 0.0199
beta_pos_strong	2	90	390	0.9171	0.918	0.052	0.0277 / 0.0172
beta_pos_strong	3	90	390	0.8810	0.885	0.048	0.0292 / 0.0194
beta_pos_strong	4	90	390	0.8590	0.866	0.045	0.0292 / 0.0195
beta_pos_weak	0	100	460	0.9491	0.949	0.251	0.0373 / 0.0216
beta_pos_weak	1	110	480	0.9258	0.932	0.109	0.0383 / 0.0229
beta_pos_weak	2	110	470	0.9483	0.945	0.191	0.0354 / 0.0209
beta_pos_weak	3	100	450	0.9388	0.935	0.151	0.0385 / 0.0232
beta_pos_weak	4	110	480	0.9258	0.931	0.131	0.0385 / 0.0240
beta_zero	0	120	820	1.0000	1.000	1.000	0.3441 / 0.0532
beta_zero	1	130	820	1.0000	1.000	1.000	0.3671 / 0.0811
beta_zero	2	120	810	1.0000	1.000	1.000	0.3436 / 0.0512
beta_zero	3	120	750	1.0000	1.000	1.000	0.3594 / 0.0671
beta_zero	4	130	800	1.0000	1.000	1.000	0.3601 / 0.0779
gnmatch_neg_strong	0	120	870	0.9991	1.000	1.000	0.7374 / 0.1486
gnmatch_neg_strong	1	130	860	1.0000	1.000	0.999	0.7294 / 0.1241
gnmatch_neg_strong	2	120	890	1.0000	0.999	0.997	0.7374 / 0.1168
gnmatch_neg_strong	3	120	800	1.0000	0.999	0.998	0.7399 / 0.1365
gnmatch_neg_strong	4	130	860	1.0000	1.000	0.998	0.7309 / 0.1107
gnmatch_pos_strong	0	120	870	1.0000	0.999	0.998	0.7446 / 0.1430
gnmatch_pos_strong	1	120	880	1.0000	1.000	1.000	0.7164 / 0.1169
gnmatch_pos_strong	2	120	890	0.9997	1.000	1.000	0.7437 / 0.1404
gnmatch_pos_strong	3	120	790	1.0000	1.000	0.999	0.7372 / 0.1230
gnmatch_pos_strong	4	130	850	1.0000	1.000	0.997	0.7293 / 0.1137

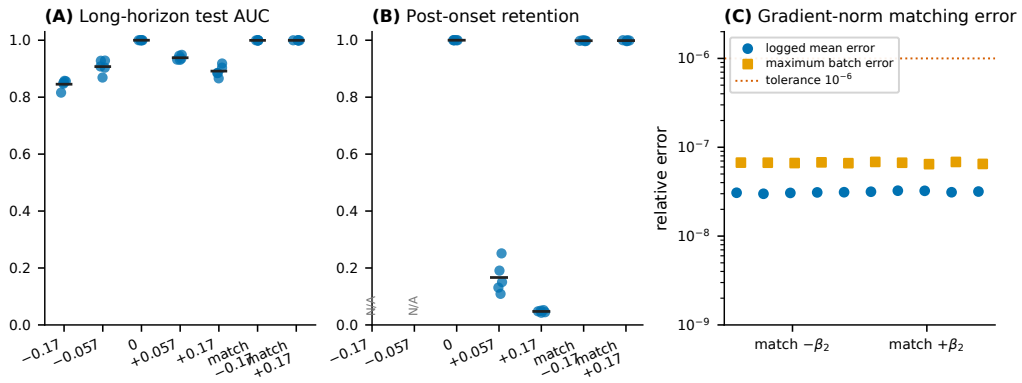

Figure 9: Long-horizon AUC, retention, and corrected matched-control errors. Binary grok-to-ungrok counts should be interpreted with duration and endpoints: matched controls show brief dips but near-unit AUC and retention, unlike positive penalties.

C.3 SECOND TASK AND DETECTOR TRANSFER

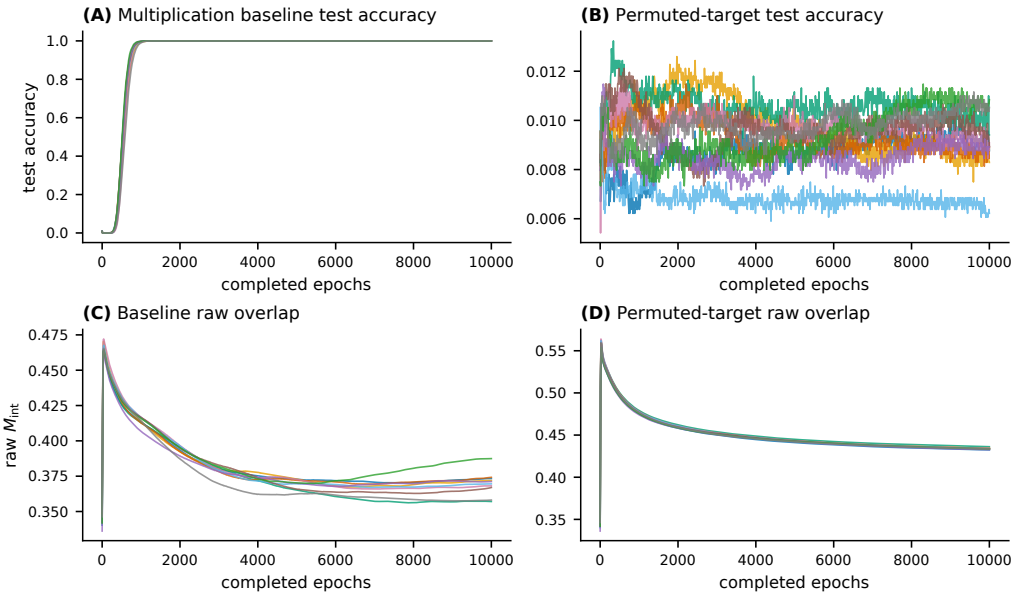

Figure 10: All modular-multiplication formal trajectories. Baseline runs robustly grok; target-permutation runs memorise training targets and remain near chance on true labels.

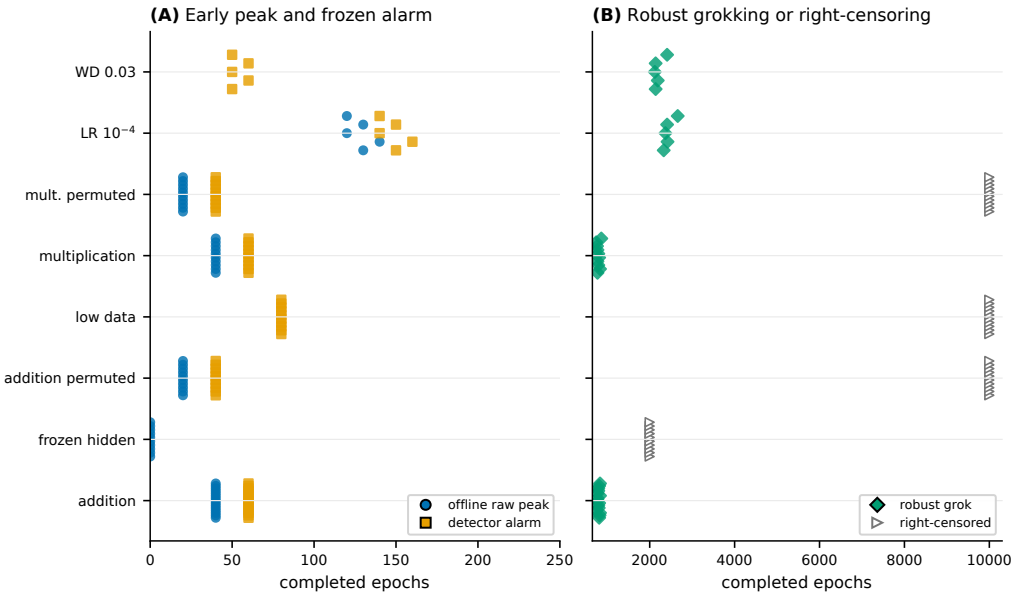

Figure 11: All frozen-detector transfer timelines. Arrows denote right-censoring.

Table 6: P0-2 modular-multiplication per-run outcomes.

Condition	Seed	t_{train}	$t_{\text{grok}}^{\text{robust}}$	Final	t_{peak}	Final raw / centred
baseline	20	120	770	1.0000	40	0.3733 / 0.0728
baseline	21	120	830	1.0000	40	0.3719 / 0.0754
baseline	22	120	790	1.0000	40	0.3571 / 0.0656
baseline	23	110	770	1.0000	40	0.3742 / 0.0737
baseline	24	120	810	1.0000	40	0.3682 / 0.0692
baseline	25	110	790	1.0000	40	0.3696 / 0.0739
baseline	26	110	770	1.0000	40	0.3670 / 0.0660
baseline	27	110	760	1.0000	40	0.3712 / 0.0821
baseline	28	120	740	1.0000	40	0.3874 / 0.0906
baseline	29	110	860	1.0000	40	0.3581 / 0.0691
shuffled_labels	20	200	cens.	0.0088	20	0.4324 / 0.0553
shuffled_labels	21	200	cens.	0.0083	20	0.4343 / 0.0553
shuffled_labels	22	200	cens.	0.0094	20	0.4362 / 0.0552
shuffled_labels	23	200	cens.	0.0089	20	0.4342 / 0.0552
shuffled_labels	24	210	cens.	0.0091	20	0.4342 / 0.0552
shuffled_labels	25	200	cens.	0.0062	20	0.4338 / 0.0553
shuffled_labels	26	200	cens.	0.0089	20	0.4335 / 0.0551
shuffled_labels	27	200	cens.	0.0089	20	0.4327 / 0.0551
shuffled_labels	28	200	cens.	0.0108	20	0.4337 / 0.0550
shuffled_labels	29	200	cens.	0.0105	20	0.4336 / 0.0551

C.4 Local randomness sensitivity

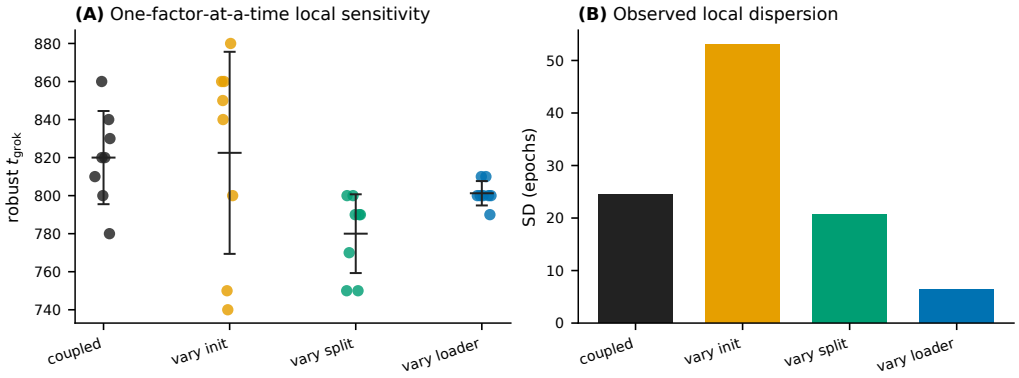

Figure 12: P0-4 local one-factor-at-a-time sensitivity around one fixed anchor. Initialisation produces the largest observed local dispersion; this is not a global variance decomposition.

Table 7: P0-4 one-factor-at-a-time local-sensitivity outcomes.

Condition	Seed	$t_{\mathrm{grok}}^{\mathrm{robust}}$	t_{peak}	Final raw / centred
coupled_reference_4090	0	840	40	0.3858 / 0.1018
coupled_reference_4090	1	860	40	0.3609 / 0.0646
coupled_reference_4090	2	820	40	0.3602 / 0.0710
coupled_reference_4090	3	810	40	0.3846 / 0.1064
coupled_reference_4090	4	820	40	0.3616 / 0.0640
coupled_reference_4090	5	780	40	0.3740 / 0.0927
coupled_reference_4090	6	830	40	0.3782 / 0.0993
coupled_reference_4090	7	800	40	0.3718 / 0.0847
vary_init	0	840	40	0.3518 / 0.0589
vary_init	1	860	40	0.3622 / 0.0688
vary_init	2	750	40	0.3778 / 0.0960
vary_init	3	880	40	0.3770 / 0.0960
vary_init	4	800	40	0.3636 / 0.0616
vary_init	5	860	40	0.3558 / 0.0630
vary_init	6	740	40	0.3587 / 0.0598
vary_init	7	850	40	0.3810 / 0.1056
vary_loader	0	810	40	0.3742 / 0.0841
vary_loader	1	800	40	0.3644 / 0.0776
vary_loader	2	810	40	0.3773 / 0.0976
vary_loader	3	800	40	0.3854 / 0.1076
vary_loader	4	790	40	0.3905 / 0.1185
vary_loader	5	800	40	0.3754 / 0.0964
vary_loader	6	800	40	0.3714 / 0.0768
vary_loader	7	800	40	0.3820 / 0.0980
vary_split	0	790	40	0.3686 / 0.0833
vary_split	1	750	40	0.3596 / 0.0632
vary_split	2	750	40	0.3809 / 0.1056
vary_split	3	770	40	0.3740 / 0.0933
vary_split	4	790	40	0.3607 / 0.0579
vary_split	5	800	40	0.3736 / 0.0862
vary_split	6	800	40	0.3689 / 0.0732
vary_split	7	790	40	0.3643 / 0.0622

D TRANSFORMER PROBE: A MEASURED ARCHITECTURE BOUNDARY

The observational results of Section 5 are scoped to one two-linear-layer ReLU MLP family, and Section 6 states that definition-level applicability does not by itself establish shared dynamics across architectures. Whether the central temporal feature of Section 5—an interior early raw-overlap peak—is a property of the observable or of the architecture is exactly the kind of boundary this paper aims to measure rather than assume. We therefore ran a preregistered, separately frozen probe in a one-layer Transformer: explicit Q/K/V causal attention with four heads, $d_{\mathrm{model}} = 128$, $d_{\mathrm{mlp}} = 512$ (227,313 parameters), ReLU activations, no LayerNorm, and no dropout. The task (modular addition mod 113, 50% split) and optimisation (AdamW, learning rate 3×10^{-4}, weight decay 0.1, batch size 64, 10,000 epochs, logging every 10 completed epochs) match the MLP protocol, but the probe is a separate protocol with its own codebase and execution environment, ten seeds per condition, and one paired permuted-target run per baseline. Configuration selection read only training and test accuracies before the formal configuration was frozen under a recorded SHA-256 hash; paired runs share split, initialisation, and loader order, differing only in the permuted optimisation targets. Full protocol, pilot, and pairing records are provided in the supplement.

Definition-level applicability and integrity. At every logging checkpoint the probe computes raw and centred M_{int} at three preregistered sites of the final token: the 512-dimensional post-ReLU feed-forward activation (the primary, MLP-analogue site), the residual stream, and the attention output. All 20 formal runs pass the logging-grid, numerical-range, active-feature, and pairing audits, and the maximum participation-ratio identity discrepancy is 1.25×10^{-12}, within the prespecified 10^{-10} tolerance. Applicability of the definition is a statement about arithmetic, not about dynamics: M_{int} remains well defined on Transformer feature matrices, and nothing follows from this alone about shared temporal structure.

Table 8: Transformer probe per-run outcomes (protocol B; separate from the 142 MLP runs of protocol A). Raw primary-site (feed-forward) t_{peak} equals 0 with boundary argmax in every run of both conditions. Permuted-target runs are right-censored with no robust threshold event; final true-label test accuracy is at chance ($1/113 = 0.00885$). Retention is the post-onset fraction of logged points at or above 0.95.

Seed	t_{train}	raw t_{peak}	boundary	robust t_{grok}	retention	perm. final test
0	70	0	yes	320	0.997	0.0077
1	180	0	yes	940	0.999	0.0096
2	130	0	yes	800	0.998	0.0096
3	150	0	yes	710	0.998	0.0072
4	120	0	yes	420	1.000	0.0077
5	110	0	yes	680	0.999	0.0106
6	150	0	yes	940	0.999	0.0089
7	110	0	yes	660	0.999	0.0111
8	90	0	yes	630	0.994	0.0091
9	110	0	yes	940	0.999	0.0075

Grokking under the matched protocol. All ten baseline runs grok (10/10; exact 95% interval $[0.69, 1.0]$): $t_{\mathrm{train}} = 122.0 \pm 31.9$ epochs and robust $t_{\mathrm{grok}} = 704.0 \pm 214.1$ epochs (range 320–940; Table 8), with a train-to-grok delay of at least 250 epochs in every run and post-onset retention at or above 0.994 (mean $\pm$ sample SD throughout). Epoch counts are not compared across the two protocols, which differ in codebase and environment.

Permuted-target non-sufficiency. The ten paired permuted-target runs fit their permuted training targets to accuracy 1.0 while true-label test accuracy remains at chance: 0/10 robust threshold events, all right-censored rather than imputed at the horizon, with final test accuracy 0.00890 ± 0.00136 against chance $1/113 = 0.00885$ (Table 8; Figure 13). Raw primary-site overlap declines throughout these runs, from 0.62–0.66 at the untrained checkpoint to 0.27–0.29 at the horizon. The non-sufficiency boundary established by the MLP controls in Section 5—overlap change without target-relevant generalization—is thus reproduced in this architecture.

The primary site does not reproduce the interior early peak. In every one of the 20 formal runs, the full-trajectory argmax of raw M_{int} at the preregistered primary site is the untrained epoch-0 checkpoint: $t_{\mathrm{peak}} = 0$ with boundary argmax in 20/20 runs (Table 8). The baseline trajectory subsequently decreases towards an interior minimum (0.26–0.32 at epochs 140–490) before rising to 0.41–0.46 at the horizon, so the interior extremum is a minimum rather than an early peak, and the MLP-style peak shape is absent at this site. At the secondary residual site, interior maxima occur after robust grokking in 9/10 baseline runs, with the latest at epoch 9,490 of 10,000, so residual-site peak times additionally carry a horizon-censoring caveat. Per-site trajectories and peak tables, including the boundary flag for every argmax, are given in the supplement.

What transfers, and what does not. Three components survive the architecture change: the definition and numerical machinery of M_{int}, the occurrence of grokking under the matched protocol, and the permuted-target non-sufficiency boundary. The interior early raw-overlap peak does not: at the preregistered primary site it is absent in 20/20 runs. Following the wording rule fixed in the preregistered protocol before the formal runs, we report this as a measured architecture boundary, not a failed run: the previously observed peak dynamics may depend on the shallow ReLU feature convention rather than on the observable itself. The probe supports no temporal-separation claim beyond the tested MLP family, no transfer of the intervention or detector results, no comparison of absolute M_{int} levels across architectures, and no statement about large models; it covers a single tiny one-layer Transformer, one task, and ten seeds per condition.

D.1 Transformer probe (protocol B)

This subsection provides the per-run detail behind Section D. It introduces no new scientific conclusions: definition-level applicability, baseline grokking, permuted-target non-sufficiency, and the architecture boundary at the preregistered primary site are the four measured results, as reported in the main text. All values below are transcribed from the protocol-B analysis artifacts; none is read from a figure.

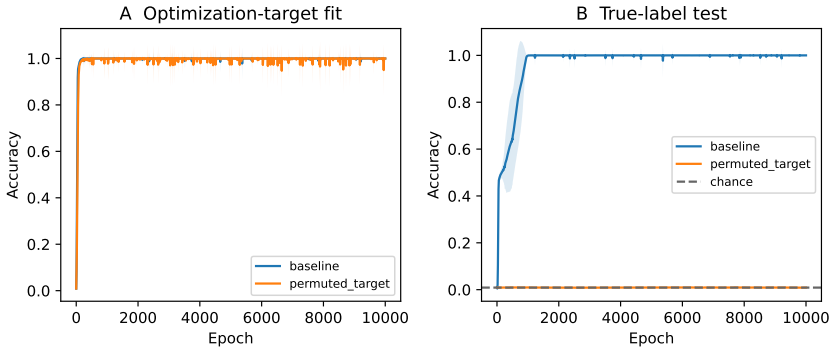

Figure 13: Transformer probe controls. (A) Optimization-target training accuracy: baseline and permuted-target runs both fit their training targets. (B) True-label test accuracy (dashed line: chance $1/113$): baseline runs grok; permuted-target runs remain at chance throughout. The probe reproduces the non-sufficiency boundary (overlap declines while true-label accuracy stays at chance) but not the primary-site interior early peak (Table 8).

Table 9: Protocol-B per-run outcomes (all 20 formal runs). `peak_at_boundary` flags whether the full-trajectory argmax of raw primary-site M_{int} falls on the untrained epoch-0 or terminal-horizon boundary. Permuted-target runs are right-censored: robust t_{grok} remains missing rather than being replaced by the horizon. Retention is the post-onset fraction of logged points at or above 0.95. Interval endpoints elsewhere are min–max over the ten runs, not mean ± SD.

Condition	Seed	t_{train}	t_{peak}	boundary	robust t_{grok}	retention	final test
baseline	0	70	0	yes	320	0.997	1.000
baseline	1	180	0	yes	940	0.999	1.000
baseline	2	130	0	yes	800	0.998	1.000
baseline	3	150	0	yes	710	0.998	1.000
baseline	4	120	0	yes	420	1.000	1.000
baseline	5	110	0	yes	680	0.999	1.000
baseline	6	150	0	yes	940	0.999	1.000
baseline	7	110	0	yes	660	0.999	1.000
baseline	8	90	0	yes	630	0.994	1.000
baseline	9	110	0	yes	940	0.999	1.000
permuted_target	0	180	0	yes	censored	–	0.0077
permuted_target	1	160	0	yes	censored	–	0.0096
permuted_target	2	180	0	yes	censored	–	0.0096
permuted_target	3	130	0	yes	censored	–	0.0072
permuted_target	4	150	0	yes	censored	–	0.0077
permuted_target	5	180	0	yes	censored	–	0.0106
permuted_target	6	230	0	yes	censored	–	0.0089
permuted_target	7	170	0	yes	censored	–	0.0111
permuted_target	8	190	0	yes	censored	–	0.0091
permuted_target	9	140	0	yes	censored	–	0.0075

Permuted-target overlap decline. In the ten permuted-target runs, raw primary-site M_{int} shows an overall decline across the observed range from 0.62–0.66 at the untrained checkpoint to 0.27–0.29 at the horizon (min–max over the ten runs; per-run endpoints in the `history.csv` files listed in Appendix F), with no interior maximum in any run, while all ten runs fit the permuted training targets to accuracy 1.0 and final true-label test accuracy remains at 0.0072–0.0111 against chance $1/113 = 0.00885$. This is the per-run basis of the non-sufficiency statement in Section D.

Number provenance. Every protocol-B value in the main text and this appendix is transcribed from `per_run_metrics.csv` (events, peaks, boundary flags, retention, final accuracies), `condition_summary.csv` (grok rates and exact binomial intervals), `pair_hash_checks.csv` (pairing audits), `integrity_summary.json` (integrity audit), `pilot_selection.json` (preregistration), the frozen `formal_config.json` and its hash sidecar, per-run

Table 10: Raw full-trajectory argmax epochs by site and condition. The primary site (`mlp_post`) has boundary argmaxes in 20/20 runs. `resid_post` is a secondary site interpreted only in the fixed learned basis; its baseline argmaxes are interior in 9/10 runs (seed 8 is the epoch-0 boundary case) but occur after robust grokking, and the latest (seed 0, epoch 9,490 of 10,000) is close enough to the horizon that these peak times carry a horizon-censoring caveat. `attn_out` is exploratory. No temporal-separation claim is made from any column of this table.

Site / condition	argmax epochs (seeds 0–9)	boundary count
`mlp_post` / baseline	0, 0, 0, 0, 0, 0, 0, 0, 0, 0	10/10
`mlp_post` / permuted	0, 0, 0, 0, 0, 0, 0, 0, 0, 0	10/10
`resid_post` / baseline	9490, 6790, 1680, 2960, 5900, 5620, 4810, 5770, 0, 5440	1/10
`resid_post` / permuted	0, 0, 0, 0, 0, 0, 0, 0, 0, 0	10/10
`attn_out` / baseline	610, 0, 0, 0, 1210, 1510, 0, 840, 1900, 1920	4/10
`attn_out` / permuted	0, 0, 0, 0, 0, 0, 0, 0, 0, 0	10/10

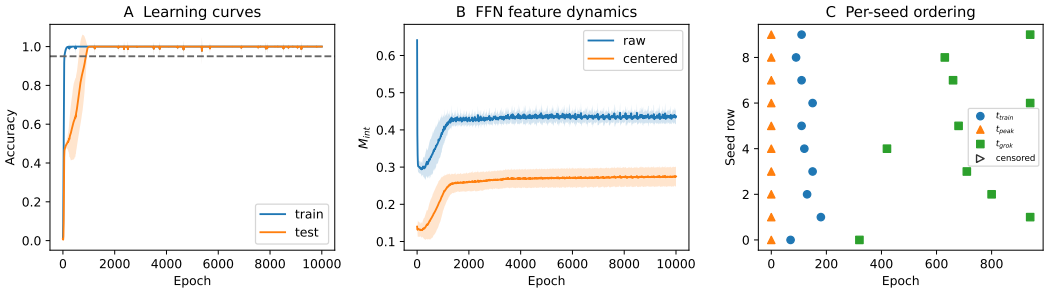

Figure 14: Protocol-B main summary (baseline). (A) Learning curves, mean ± sample SD, with the 0.95 threshold. (B) Primary-site (`mlp_post`) raw and centred M_{int}. (C) Per-seed event ordering: t_{peak} markers are the epoch-0 boundary argmaxes in every run, not interior peaks; hollow markers denote right-censored t_{grok}, which remains missing. The interior early peak observed in protocol A is absent at this primary site (Section D).

`metadata.json` files, or the per-run `history.csv` files. No numerical value is inferred from a plotted image, and no protocol-B statistic is pooled with the 142 protocol-A runs.

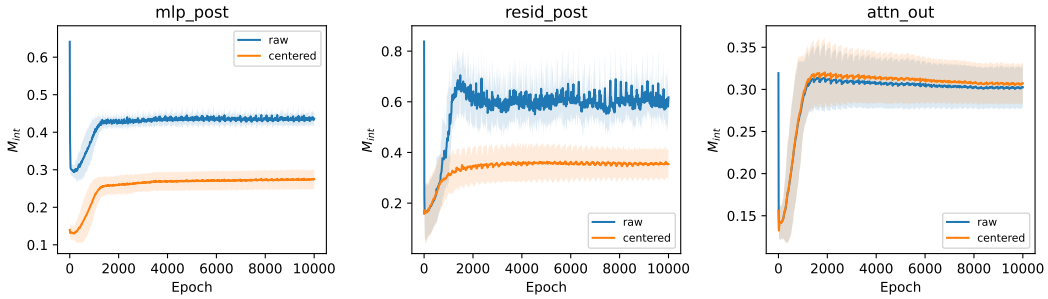

Figure 15: Protocol-B feature-site dynamics (baseline, mean ± sample SD), raw and centred conventions. `mlp_post` is the preregistered primary site. `resid_post` is shown only in the fixed learned basis; its raw maxima occur after grokking in 9/10 runs and are subject to the horizon-censoring caveat of Table 10. `attn_out` is exploratory. Absolute M_{int} levels are not comparable across sites or architectures.

E Reproducibility, Integrity, and Artifact Notes

The evidence-building script inventories exactly 142 formal runs: 45 first-stage baseline/control runs, 35 P0-1 runs, 20 P0-2 runs, 10 P0-3 shift runs, and 32 P0-4 runs. It reconstructs every reported endpoint directly from `history.csv`. The compact release contains the derived per-run table, condition summaries, detector transfer rows, run inventory, compressed trajectories, plotting scripts, and LaTeX table fragments. Original training archives remain the authoritative source.

Every formal history uses a regular epoch grid. Across the formal set, all norm-conditioned energy checks are true, the maximum PR-identity discrepancy is 1.71×10^{-13}, and the minimum active-feature count is 361. The P0-1 signed conditions are paired by split, initial-parameter, and loader-prefix hashes. Corrected v1.1 matched runs replace the superseded matched runs. P0-2 baseline and target-permutation runs are similarly paired. All 75 detector transfer rows use one byte-identical frozen detector configuration. P0-4 hash checks validate the one-factor design.

The first-stage environment is recorded as Python 3.9.18, PyTorch 1.12.1, NumPy 1.21.5, and CPU execution. P0-4 includes Python, PyTorch, CUDA, driver, GPU, and package snapshots. The P0-1–3 archive does not contain an equally complete server-side environment snapshot or pytest execution transcript; this is a release blocker to be filled without retraining. The source includes a beta-zero first-step equivalence test, but the final anonymous artifact should add the actual execution log. Target-permutation arrays should be released or accompanied by reproducible hashes.

Protocol B integrity. The protocol-B analysis enforces, and the artifact records, the following: 20/20 formal runs with status `complete`; a regular logging grid $0, 10, \ldots, 10{,}000$ in every run; accuracy and M_{int} ranges and $K_{\text{active}} \geq 2$ checks passing at every logged point; a maximum participation-ratio identity discrepancy of 1.25×10^{-12} against the prespecified 10^{-10} tolerance across all sites and conventions; and paired-control hash audits passing 10/10 (config hash, seed bundle, split hash, true-targets hash, initial-state hash, and loader-prefix hash match within each seed pair; only the optimisation-target hash differs). Right-censored robust t_{grok} values remain missing in the analysis outputs.

Protocol separation. The 142-run inventory in the first paragraph of this appendix refers exclusively to protocol A. Protocol B is a separate frozen protocol with its own codebase, configuration, seed bundle, and execution environment, and no protocol-B statistic is pooled with, or averaged against, any protocol-A statistic anywhere in this paper. The protocol-A detector SHA above was never evaluated on protocol-B runs; the from-init decline of the Transformer's raw overlap would make an MLP-calibrated alarm trivially fire, so any cross-protocol detector statement would require a preregistered recalibration study, which is future work. Protocol B ships its complete execution logs (`logs/formal_*.log`) and analysis manifests; this does not alter the protocol-A release blocker disclosed above.

No value in the manuscript is read from a plotted image. The frozen build order is raw history/JSONL $>$ summary/config/manifest $>$ derived per-run metrics $>$ aggregate tables $>$ figures and prose. Bootstrap analyses use 20,000 run-level resamples with seed 20260803. Because P0-1 has five seeds per condition, inference emphasises paired points, effect size, bootstrap intervals, censoring, AUC, retention, and final outcomes rather than uncorrected p-values.

F EVIDENCE MANIFEST

The package intentionally excludes the 184 MB raw archives from the Overleaf project. They should be distributed as an anonymous supplementary artifact, with archive SHA-256 hashes added to this section before submission.

Table 11: Evidence manifest of the compact release.

Artifact	Role
evidence/per_run_metrics_142.csv	All formal-run derived endpoints, censoring, overlap endpoints, integrity checks.
evidence/run_inventory_142.csv	Unique formal-run inventory and source history paths.
evidence/p0_1_summary.csv	Intervention dose, robust onset, AUC, retention, endpoints, match tolerance.
evidence/p0_2_summary.csv	Second-task baseline and target-permutation summaries.
evidence/p0_3_detector_summary.csv	Alarm versus robust-grok outcomes by condition and subset.
evidence/p0_4_summary.csv	Local dispersion, robust summaries, bootstrap intervals.
evidence/claim_evidence_matrix.csv	Safe claim wording, evidence role, caveat, manuscript placement.
evidence/claim_source_rows.csv	Representative claims mapped to logical raw-history paths and CSV line ranges.
data/trajectories_142.csv.gz	Compact machine-readable trajectories used to regenerate figures.
scripts/build_evidence.py	Frozen endpoint reconstruction and integrity audit.
scripts/make_figures.py	Main and supplementary figure generation.
scripts/make_tables.py	Main and supplementary LaTeX table generation.
scripts/audit_release.py	Independent cross-file consistency audit for the compact release evidence.
evidence/revision_audit.json	Machine-readable release-audit results and mismatch lists.
REVISION_AUDIT.md	Human-readable release-audit summary and scope boundary.
RAW_ARCHIVE_SHA256.txt	SHA-256 digests of the two authoritative source archives.
LOCAL_PYTEST_CURRENT_ENV.txt	Independent 26-test execution in the assembly environment; not a replacement for the original server transcript.
Protocol B (Transformer probe; separate frozen protocol; 20 formal runs; never pooled with protocol A)	
results/formal/T0/analysis/per_run_metrics.csv	Per-run events, peaks with boundary flags, censoring, retention, and final accuracies for the 20 protocol-B runs.
results/formal/T0/analysis/condition_summary.csv	Grok rates, exact binomial intervals, and peak summaries by condition (protocol B).
results/formal/T0/analysis/pair_hash_checks.csv	Paired-control hash audits (10/10 pass; protocol B).
results/formal/T0/analysis/integrity_summary.json	Run-count, logging-grid, range, K_{active}, and PR-identity audit results (protocol B).
results/formal/T0/analysis/audit_report.md	Human-readable protocol-B formal audit summary.
results/protocol/formal_config.json + .sha256	Frozen protocol-B configuration and canonical SHA-256 hash.
results/protocol/pilot_selection.json	Behaviour-only pilot selection record (selection_reads_M_int = false).
results/formal/T0/baseline/seed_*/history.csv and results/formal/T0/permuted_target/seed_*/history.csv	Raw per-run logged histories; authoritative source for every protocol-B number.

Table 11: Evidence manifest of the compact release (continued).

Artifact	Role
results/formal/T0/analysis/figure_transformer_*.pdf	Supplementary protocol-B figures (main, control, feature sites).
mint_xarch/, tests/, VERSION.txt	Protocol-B source code, unit tests, and version (1.0.2).
docs/Mint_Transformer_PurePython_PyTorch_Final_Manual_CN.pdf	Preregistered protocol-B manual: hypotheses H1–H4, site grading, boundary-peak disclosure mandate, and the wording ladder fixed before the formal runs.