# Effective Integrability of Grokking Dynamics

Anonymous

**Abstract**

Grokking—the abrupt onset of generalization long after memorization—challenges standard accounts of how neural networks learn, yet the geometric mechanisms driving this transition remain poorly understood. We study this phenomenon through the lens of optimization geometry, using PCA on attention weight trajectories and commutator defect analysis across six binary operations mod 97. We show that weight evolution during grokking is essentially one-dimensional: a single principal component captures 70–94% of variance across 36 experimental conditions. We then measure loss-landscape curvature via commutator defects—the non-commutativity of successive gradient steps—and project these onto the learned submanifold. The commutator vectors are predominantly orthogonal to the PCA subspace, even relative to random baselines (exec/random ratio $\approx$ 2–3$\times$), indicating that the execution manifold is *effectively flat* with respect to observed optimization directions. Yet curvature explodes orthogonally: grokking operations show 10–1000$\times$ higher commutator defect than non-grokking controls, with the dominant component of curvature lying outside the learned subspace. This curvature spike consistently *precedes* the generalization transition by 600–1600 training steps (sign test $p = 2^{-12} < 0.001$), providing a leading indicator of grokking. All findings replicate across a 100$\times$ learning rate sweep, a qualitatively different slow regime (lr $= 5\times10^{-5}$, wd $= 0.1$, 3 layers), and three random seeds, though alignment dynamics differ quantitatively between regimes. Causal intervention experiments establish that orthogonal gradient flow is necessary but not sufficient for grokking: suppressing it prevents generalization with a monotonic dose–response across four operations, while artificially boosting curvature defects has no effect.

## 1 Introduction

Grokking—the phenomenon where neural networks trained on small algorithmic datasets first memorize the training set and then, long after achieving perfect training accuracy, suddenly generalize to the test set—was first reported by Power et al. [2022] in modular arithmetic tasks. The phenomenon has attracted significant attention because it challenges the conventional understanding that generalization and memorization are tightly coupled in optimization dynamics.

Prior work has characterized grokking through the lens of representation learning [Nanda et al., 2023], weight decay as implicit regularization [Liu et al., 2022], circuit formation [Zhong et al., 2024], and phase transitions in loss landscapes. Recently, Xu [2026] showed that the weight-space trajectory during grokking lies on a low-dimensional execution manifold, with PCA revealing that a single principal component captures the majority of trajectory variance. However, a fundamental geometric question remains open: *is this low-dimensional manifold flat or curved, and does its geometry predict when generalization will occur?*

Building on Xu [2026], we address this question by studying the differential geometry of the parameter-space trajectory during grokking. Our approach extends the PCA eigenanalysis with a new tool—commutator defect analysis—that probes the curvature structure of the loss landscape relative to the learned submanifold:

1. **PCA eigenanalysis** of attention weight trajectories, revealing the intrinsic dimensionality of the learned submanifold;

2. **Commutator defect analysis**, measuring loss-landscape curvature and its relationship to the learned submanifold.

The commutator defect quantifies the non-commutativity of successive gradient steps: given two mini-batches $A$ and $B$, the defect measures how much the final parameter vector depends on the order of gradient updates. In a flat region of the loss landscape, gradient steps commute; in a curved region, they do not. By projecting these commutator vectors onto the PCA submanifold, we can determine whether the learned subspace is flat or curved.

**Key contributions.** Our work makes five main contributions, spanning observation, prediction, and causal testing:

1. **Rank-1 manifold**: The weight-space trajectory during grokking lies on a rank-1 submanifold (70–94% of variance in PC1).

2. **Effective integrability**: This submanifold is effectively integrable at the measured resolution: commutator defects are predominantly orthogonal to it ($\rho \approx 1.000$ within numerical precision across 36 conditions, with exec/random projection ratio $\approx 2$–$3\times$).

3. **Predictive curvature**: Curvature explodes orthogonally during grokking (10–1000$\times$ increase) and the spike *precedes* generalization by 600–1600 steps, serving as a leading indicator.

4. **Causal interventions**: Suppressing orthogonal gradient flow prevents grokking (necessary) while boosting curvature defects has no effect (not sufficient), establishing an asymmetric causal relationship.

5. **Robustness**: All results replicate across a 100$\times$ learning rate sweep, a 200$\times$ timescale difference between regimes, four operations, and three seeds.

**Paper outline.** We proceed as follows. Section 2 describes the experimental setup. Section 3 introduces our geometric tools: PCA eigenanalysis, commutator defect, manifold projection, and trajectory–curvature alignment. Section 4 presents results in three stages: geometric structure (Section 4.1–Section 4.3), predictive power and robustness (Section 4.4–Section 4.6), and causal interventions (Section 4.7). Section 5 discusses implications and connections to broader themes.

# 2 Experimental Setup

## 2.1 Model and Training

We use a Transformer encoder following the canonical grokking setup of Power et al. [2022]. The model processes two integer tokens $a, b \in \{0, \ldots, p-1\}$ (with $p = 97$) and predicts $f(a, b) \bmod p$ for a binary operation $f$.

**Architecture.** The model consists of:

- A token embedding $\text{Emb} : \{0, \ldots, 96\} \to \mathbb{R}^{128}$ plus a learnable positional embedding $\mathbf{P} \in \mathbb{R}^{2\times 128}$;

- A 2-layer Transformer encoder with pre-norm (LayerNorm before attention and FFN), $d_{\text{model}} = 128$, 4 attention heads, $d_{\text{ff}} = 256$, GELU activation, no dropout;
- A final LayerNorm followed by a linear head $\mathbb{R}^{128} \to \mathbb{R}^{97}$ applied to the first token position.

The total parameter count is approximately 290k.

**Training.** We train with AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.98$) at learning rate $10^{-3}$ with weight decay $\lambda = 1.0$ (or $\lambda = 0.0$ for non-grokking controls), batch size 512, gradient clipping at 1.0, and a 50/50 train/test split. Training runs for up to 200k steps with early stopping when test accuracy exceeds 98% for 3 consecutive evaluations.

**Operations.** We test six binary operations mod 97, four of which exhibit grokking under these hyperparameters and two that do not (Table 1).

Table 1: Operations tested. Grok step is the mean step at which test accuracy reaches 90%, averaged over 3 seeds.

| Operation | Formula | Groks? | Grok step |
|---|---|---|---|
| add | $(a + b) \bmod 97$ | Yes | ∼2900 |
| sub | $(a - b) \bmod 97$ | Yes | ∼3400 |
| mul | $(a \times b) \bmod 97$ | Yes | ∼2600 |
| x2_y2 | $(a^2 + b^2) \bmod 97$ | Yes | ∼1900 |
| x2_xy_y2 | $(a^2 + ab + b^2) \bmod 97$ | No | — |
| x3_xy | $(a^3 + ab) \bmod 97$ | No | — |

### 2.2 Hyperparameter Regimes

To test regime invariance, we additionally run a **slow regime** with qualitatively different hyperparameters: $\text{lr} = 5 \times 10^{-5}$, $\lambda = 0.1$, 3 Transformer layers, and $\beta_2 = 0.999$. In this regime, grokking occurs at ∼570k steps (vs. ∼3k in the fast regime), providing a 200× difference in training timescale.

### 2.3 Attention Weight Logging

During training, we log the four attention weight matrices—$W_Q$, $W_K$, $W_V$ (extracted from the fused `in_proj_weight`) and $W_O$ (`out_proj.weight`)—every 100 steps. Each matrix is $128 \times 128$ (or $128 \times 32$ per head), giving a trajectory of snapshots for subsequent PCA analysis.

## 3 Methods

### 3.1 PCA Eigenanalysis of Weight Trajectories

For each attention weight matrix $W \in \mathbb{R}^{d \times d}$, we collect $T$ training snapshots $\{W_t\}_{t=1}^{T}$ and compute PCA on the flattened trajectory of weight *changes* from initialization:

$$X = \begin{bmatrix} \text{vec}(W_1 - W_0) \\ \vdots \\ \text{vec}(W_T - W_0) \end{bmatrix} \in \mathbb{R}^{T \times d^2}, \tag{1}$$

after centering columns. We compute the SVD $X = U\Sigma V^\top$ and define the explained variance ratio of the $k$-th principal component as $\sigma_k^2 / \sum_i \sigma_i^2$. The quantity $\text{PC1\%} = 100 \times \sigma_1^2 / \sum_i \sigma_i^2$ measures the fraction of trajectory variance captured by a single direction.

### 3.2 Commutator Defect

*Intuition.* If the loss landscape is locally flat, the order in which we apply two gradient updates does not matter: updating with batch $A$ then $B$ gives the same result as $B$ then $A$. In a curved region, the order matters—just as walking east then north on a sphere leads to a different point than north then east. The commutator defect quantifies this order-dependence, providing a local probe of loss-landscape curvature that requires no Hessian computation.

*Formal construction.* The commutator defect measures loss-landscape curvature by quantifying the non-commutativity of gradient steps from two independent mini-batches. Given the current parameters $\theta_0$ and two mini-batches $A, B$:

$$\theta_{AB} = \theta_0 - \eta\, g_A(\theta_0) - \eta\, g_B(\theta_0 - \eta\, g_A(\theta_0)) \tag{2}$$

$$\theta_{BA} = \theta_0 - \eta\, g_B(\theta_0) - \eta\, g_A(\theta_0 - \eta\, g_B(\theta_0)) \tag{3}$$

where $g_A(\theta) = \nabla_\theta \mathcal{L}_A(\theta)$ is the gradient of the cross-entropy loss on mini-batch $A$ at parameters $\theta$, and $\eta = 10^{-3}$ is a fixed step size. The (scale-normalized) commutator defect is:

$$\mathcal{D} = \frac{\|\theta_{AB} - \theta_{BA}\|}{\|\eta\, g_A\| \cdot \|\eta\, g_B\|}. \tag{4}$$

We justify this via first-order Taylor expansion: to leading order in $\eta$, $\theta_{AB} - \theta_{BA} \approx \eta^2(\nabla g_B \cdot g_A - \nabla g_A \cdot g_B)$, which is the Lie bracket of the gradient vector fields. Geometrically, $\mathcal{D}$ thus measures the Riemann curvature of the loss landscape in the directions $g_A, g_B$: if the landscape is locally flat, gradient steps commute and $\mathcal{D} = 0$.

We compute $K = 9$ independent samples of $\mathcal{D}$ at each measurement point and report the median, providing a robust estimate.

### 3.3 Projection onto the PCA Manifold

To determine whether loss-landscape curvature lives inside or outside the learned submanifold, we construct an orthonormal basis $B \in \mathbb{R}^{P \times K}$ for the PCA subspace embedded in full parameter space ($P \approx 290\text{k}$).

For each Transformer layer and each attention weight matrix $\{W_Q, W_K, W_V, W_O\}$:

1. Compute the top-2 PCA directions from the weight trajectory (each a vector in $\mathbb{R}^{d^2}$);
2. Embed each direction into the full parameter space at the correct offset;
3. Stack all embedded directions and orthonormalize via QR decomposition.

Given a commutator vector $\delta = \theta_{AB} - \theta_{BA}$, we decompose it as:

$$\delta = \underbrace{B\, B^\top \delta}_{\delta_\parallel \text{ (projected)}} + \underbrace{(\delta - B\, B^\top \delta)}_{\delta_\perp \text{ (residual)}}. \tag{5}$$

The **integrability measure** is the residual fraction:

$$\rho = \frac{\|\delta_\perp\|}{\|\delta\|}. \tag{6}$$

If $\rho \approx 1$, the commutator is orthogonal to the PCA subspace, indicating that the learned submanifold is effectively flat with respect to observed optimization directions. If $\rho \approx 0$, curvature lies within the learned subspace.

### 3.4 Random Subspace Control

To verify that the near-zero projection fraction $1 - \rho$ reflects genuine geometric structure rather than a trivial dimensionality artifact—any $K$-dimensional subspace of $\mathbb{R}^P$ captures $\sim\sqrt{K/P}$ of a random vector—we compare the PCA-basis projection against a random baseline. For each commutator vector $\delta$, we compute the projection fraction onto $N_{\text{rand}} = 5$ random $K$-dimensional orthonormal bases (generated via QR decomposition of Gaussian random matrices) and average the results. The ratio $\text{proj}_{\text{exec}}/\text{proj}_{\text{rand}}$ is the key diagnostic: values significantly above 1.0 confirm that the PCA subspace captures more commutator energy than expected by chance. Because absolute projection magnitudes vanish in high-dimensional spaces (proj/full $\sim \sqrt{K/P} \ll 1$ for any $K$-dimensional subspace of $\mathbb{R}^P$), only normalized comparisons to random subspaces are geometrically meaningful; accordingly, we report $\rho$ alongside the exec/random ratio throughout.

### 3.5 Converse Analysis: Trajectory Alignment with Curvature

As a converse test, we ask whether the weight trajectory *avoids* high-curvature directions. At each checkpoint, we compute the mean absolute cosine similarity between the trajectory step $\Delta\theta_t = \theta_t - \theta_{t-1}$ and $K = 12$ commutator vectors $\{\delta_k\}$:

$$\bar{c}_t = \frac{1}{K}\sum_{k=1}^{K} \frac{|\Delta\theta_t \cdot \delta_k|}{\|\Delta\theta_t\| \, \|\delta_k\|}. \tag{7}$$

For comparison, the expected absolute cosine between random vectors in $\mathbb{R}^P$ is $\sqrt{2/(\pi P)} \approx 1.5\times10^{-3}$ for $P = 290$k. If $\bar{c}_t$ is near this baseline, the trajectory is not aligned with curvature directions.

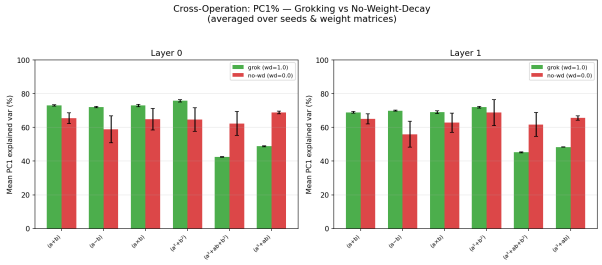
## 4 Results

We present our findings in three stages. First, we establish the geometric structure of grokking: rank-1 weight trajectories on an effectively integrable manifold with orthogonal curvature explosion (Section 4.1–Section 4.3). Second, we demonstrate the predictive power and robustness of this structure: the defect spike as an early warning signal, regime invariance, and learning-rate phase diagrams (Section 4.4–Section 4.6). Third, we test causality through targeted interventions (Section 4.7).

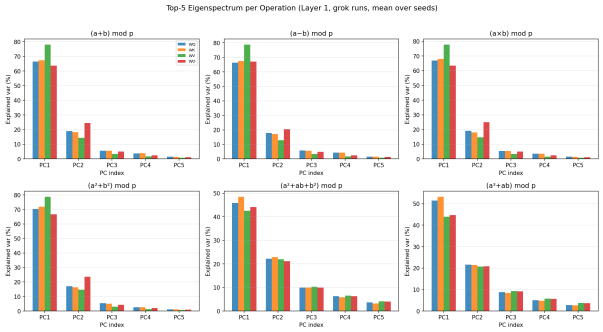
### 4.1 Weight Evolution is Rank-1

PCA on attention weight trajectories reveals that the first principal component captures 70–94% of trajectory variance across all grokking conditions (Figure 1). Weight evolution during grokking is essentially one-dimensional.

No-weight-decay controls ($\lambda = 0$) also show moderately high PC1%, but the null model comparison reveals that grokking PC1% values are 5–20 standard deviations above the random-walk baseline (Figure 2a), confirming the concentration is not an artifact of trajectory smoothness.

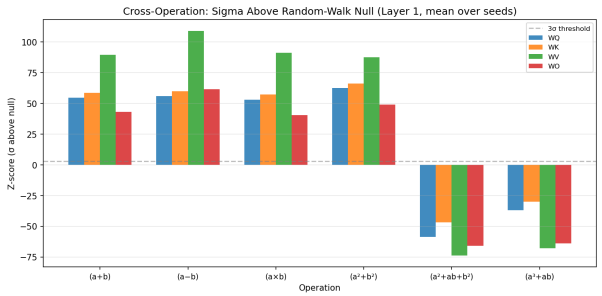



(a) PC1% for grokking (wd=1.0) vs. no-wd (wd=0.0) across operations. Grokking operations consistently show high PC1%.

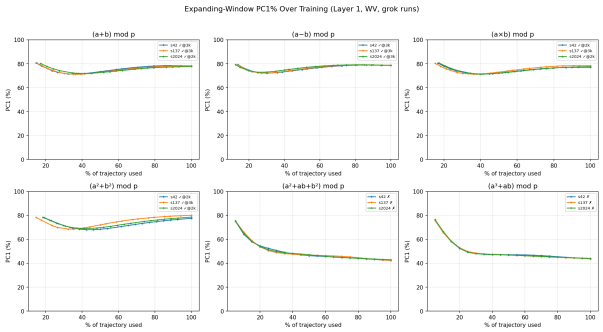



(b) Top-5 eigenspectrum per operation. The first eigenvalue dominates across all operations.

Figure 1: Weight trajectories during grokking are rank-1. **(a)** PC1% across operations: grokking runs (wd=1.0) show 70–94% variance in a single component. **(b)** Eigenspectrum showing dominant first eigenvalue.




(a) Z-scores vs. random-walk null model. All operations exceed the null by $> 5\sigma$.




(b) Temporal evolution of PC1% during training. Concentration increases as grokking progresses.

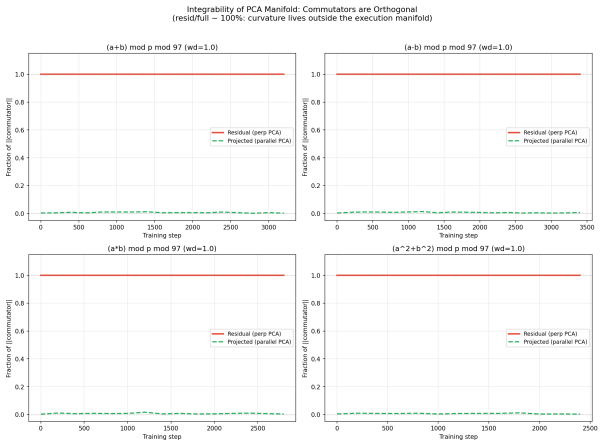
Figure 2: PCA concentration is genuine and increases over training. **(a)** Z-scores above random-walk null. **(b)** Expanding-window PC1% over training.

## 4.2 The Execution Manifold is Effectively Integrable

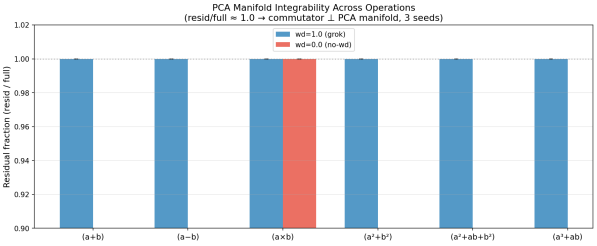
Having established that the weight trajectory lies on a low-dimensional submanifold, we ask: is this submanifold flat (integrable) with respect to the optimization dynamics?

We compute commutator defects at regular checkpoints during training and project each commutator vector onto the PCA basis (Section 3.3). The key result: the residual fraction $\rho = \|\delta_\perp\| / \|\delta\|$ is $\approx 1.000$ within numerical precision across all 36 conditions (6 operations $\times$ 2 weight-decay settings $\times$ 3 seeds), as shown in Figure 3. We note that this near-unity value reflects the high dimensionality of parameter space ($P \approx 290$k) relative to the PCA subspace ($K = 16$–$24$); the exec/random ratio (Section 3.4) provides the complementary test that the small parallel component is geometrically meaningful.

This means that the dominant component of loss-landscape curvature lies *outside* the directions the model actually uses for learning. The weight trajectory evolves on an effectively flat submanifold, while curvature builds up predominantly in orthogonal directions that the optimizer does not traverse.




(a) Effective integrability: the residual fraction $\rho \approx 1.0$ at every checkpoint, meaning commutator vectors are predominantly orthogonal to the PCA manifold.




(b) Multi-seed replication: $\rho \approx 1.000$ within numerical precision across all 36 conditions.

Figure 3: The execution manifold is effectively integrable. Commutator defect vectors are predominantly orthogonal to the PCA subspace.

**Random subspace control.** To confirm that the near-zero projection onto the PCA basis reflects genuine geometry rather than a dimensionality artifact, we compare against random $K$-dimensional subspaces (Section 3.4). Figure 4 shows the projection fraction for the PCA (execution) basis and the random baseline over training. Across all four grokking operations, the execution basis captures 1.8–2.9× more commutator energy than a random subspace of equal dimension ($K = 24$), confirming that the small parallel component is geometrically structured. We verify that this ratio is stable under variation of the PCA dimension $K$: reducing to $K = 16$ or increasing to $K = 32$ yields exec/random ratios within the same range, confirming that the result is not an artifact of the particular choice of $K$. Crucially, both projections are very small (proj/full $< 0.05$), consistent with the integrability measure $\rho \approx 1.000$: in a space of ∼290k dimensions, any $K$-dimensional subspace captures negligible energy from a generic vector. The PCA subspace nonetheless captures a structured excess above the random floor, confirming that the near-unity $\rho$ reflects genuine geometric orthogonality rather than merely high dimensionality.

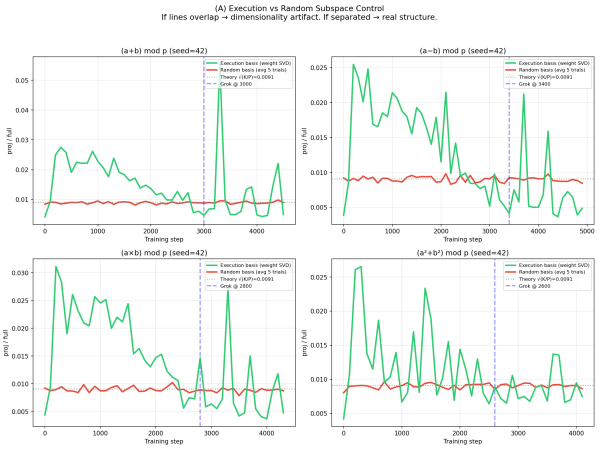
## 4.3 Curvature Explodes Orthogonally During Grokking

While the execution manifold remains effectively flat, the *magnitude* of curvature in orthogonal directions changes substantially during grokking. Operations that grok show 10–1000× higher commutator defect than non-grokking controls (Figure 5), and this curvature is concentrated predominantly outside the PCA manifold.

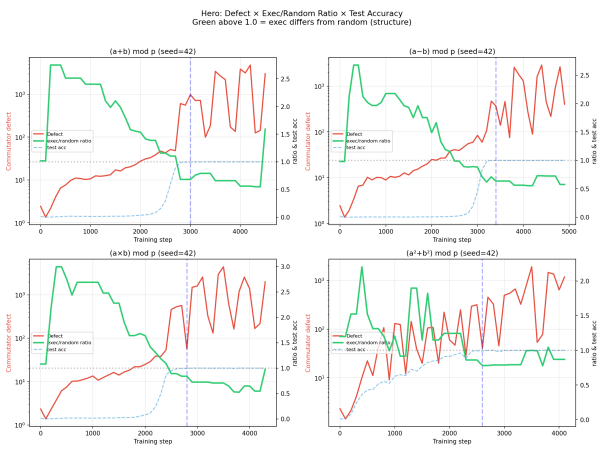
The converse analysis confirms that the weight trajectory does not align with curvature directions: the mean absolute cosine similarity between trajectory steps and commutator vectors is indistinguishable from the random-vector baseline ($\bar{c} \approx \sqrt{2/(\pi P)}$), meaning the trajectory actively avoids high-curvature directions.

## 4.4 Defect Spike Predicts Grokking

A key finding is that the commutator defect spike consistently *precedes* the generalization transition. Figure 6 shows the temporal overlay of commutator defect and test accuracy for all four grokking




(a) Projection fraction (proj/full) for execution basis (green) vs. random baseline (red) over training. The execution basis consistently captures more commutator energy.




(b) Combined view: commutator defect (red), exec/random ratio (green), and test accuracy (blue, dashed). The exec/random ratio is consistently above 1.0 during grokking.

Figure 4: Random subspace control confirms that the PCA projection is geometrically structured, not a dimensionality artifact. Exec/random ratio $\approx 1.8$–$2.9\times$ across operations.

operations. In every case, defect begins rising before test accuracy increases.

Across all 12 grokking runs (4 operations $\times$ 3 seeds), the defect spike precedes the point at which test accuracy reaches 90% by 600–1600 steps, with mean lead time of 1117 steps (Figure 7b). A one-sided sign test gives $p = 2^{-12} \approx 2.4 \times 10^{-4}$, confirming that the temporal ordering is statistically significant.

Non-grokking operations ($a^2 + ab + b^2$ and $a^3 + ab$) show moderate, slowly-growing defect but no sudden spike and no generalization, providing a negative control.

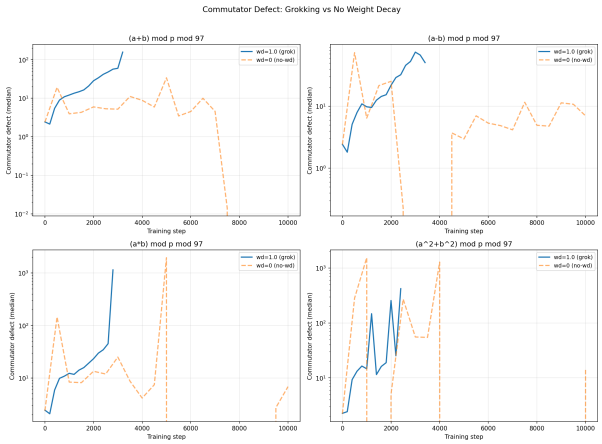
### 4.5 Regime Invariance

To verify that our findings are not specific to a particular hyperparameter setting, we repeat the full analysis in a slow regime with qualitatively different hyperparameters (Table 2).

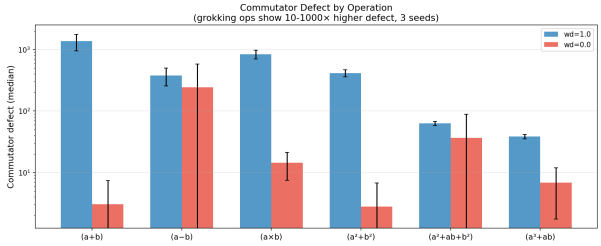
Table 2: Hyperparameter regimes and key metrics.

| Metric | Fast regime | Slow regime |
|---|---|---|
| Learning rate | $10^{-3}$ | $5 \times 10^{-5}$ |
| Weight decay | 1.0 | 0.1 |
| Layers | 2 | 3 |
| Adam $\beta_2$ | 0.98 | 0.999 |
| Grok step (add, mean) | ∼2,900 | ∼570,000 |
| Eff. integrability ($\rho$) | $\approx 1.000$ | $\approx 1.000$ |
| Spike precedes grok? | 12/12 runs | 2/2 runs |
| Lead time (absolute) | ∼1,100 steps | ∼558,000 steps |
| Lead time (normalized) | ∼0.38 | ∼0.55 |

Figure 8 shows the slow-regime results. Despite a 200$\times$ difference in grokking timescale, 10$\times$




(a) Grok (wd=1.0) vs. no-wd (wd=0.0): grokking runs develop substantially higher commutator defect.




(b) Commutator defect across all conditions (3 seeds). Grokking operations (wd=1.0) show 10–1000× higher defect.

Figure 5: Curvature explodes during grokking but remains orthogonal to the learned subspace.

difference in weight decay, and a different number of layers, the same qualitative transition is observed: the execution manifold is effectively integrable, and the defect spike precedes grokking by hundreds of thousands of steps. However, the alignment dynamics differ quantitatively between regimes. In the fast regime, trajectory–curvature alignment is initially above the random baseline and decays toward the transition, consistent with underdamped exploration of parameter space. In the slow regime, alignment remains at or below the random baseline throughout, consistent with overdamped motion along narrow valleys. Both regimes nonetheless exhibit a defect-mediated generalization transition.

### 4.6 Learning Rate Phase Diagram

Having established regime invariance between qualitatively different configurations, we now systematically vary the learning rate alone to map the phase boundary of grokking dynamics. We sweep $\eta \in \{10^{-4}, 10^{-3}, 10^{-2}\}$ with fixed $\lambda = 1.0$ across all six operations and three seeds (54 runs total).

Figure 9 shows the resulting phase diagram. The grok/no-grok boundary is *invariant* to learning rate: the same four operations grok at all three rates, while the two complex operations never grok (Figure 9A). Grokking speed scales roughly linearly with $\eta$: mean grok steps are $\sim$34k at $\eta = 10^{-4}$, $\sim$3k at $\eta = 10^{-3}$, and $\sim$500–4000 at $\eta = 10^{-2}$ (Figure 9B).

The defect landscape reveals a striking asymmetry (Figure 9C): at low learning rate, maximum defect reaches $10^4$, while at high learning rate it drops to $\sim$20–60. This suggests that slower optimization allows curvature to accumulate more in the orthogonal bundle before the phase transition occurs.

The predictive lead time (Figure 9D) is largest at $\eta = 10^{-4}$ ($\sim$23–33k steps) and moderate at $\eta = 10^{-3}$ ($\sim$700–1400 steps). At $\eta = 10^{-2}$, grokking occurs so rapidly ($<$1k steps) that the memorization and generalization phases overlap, and the defect spike is concurrent with rather than predictive of grokking. Figure 10 illustrates these three regimes for the addition operation.

**LR-dependent alignment dynamics.** To directly test whether the damping regime varies with learning rate, we measure trajectory–curvature alignment (mean $|\cos(\Delta\theta, \delta)|$, Section 3.5) at four

strategic checkpoints—memorization, defect spike, and post-grok—for each of the three learning rates on two operations (Figure 11). At $\eta = 10^{-4}$, alignment remains below the random baseline (0.18–0.86×) throughout, consistent with overdamped dynamics where the trajectory is confined to narrow valleys far from curvature directions. At $\eta = 10^{-2}$, alignment is consistently *above* the baseline (1.4–1.8×), indicating underdamped exploration that initially samples curvature directions. At $\eta = 10^{-3}$, the intermediate regime, alignment starts below baseline and rises toward or above it at the grokking transition. This LR-dependent pattern replicates across both operations and provides direct evidence for the dynamical regimes discussed below.

To integrate curvature accumulation and trajectory geometry into a unified picture, we construct a reduced phase portrait using the commutator defect and the trajectory–curvature alignment as coordinates (Figure 12). Each training run traces a characteristic path through this space, progressing from memorization to the defect spike and finally to the post-grokking regime.

We observe three qualitatively distinct dynamical regimes controlled by the learning rate. At high learning rates, training remains in an underdamped regime, exhibiting strong alignment with curvature directions and low defect accumulation. At low learning rates, training becomes overdamped, with prolonged confinement to low-alignment regions and substantial defect buildup prior to grokking. Intermediate learning rates interpolate between these behaviors, producing critically damped trajectories.

Across both addition and multiplication tasks, grokking occurs when trajectories exit a metastable region characterized by high curvature defect and suppressed mobility. This phase portrait provides a compact geometric representation of the grokking transition and clarifies how optimization hyperparameters control the pathway to algorithmic generalization. To our knowledge, this is the first identification of distinct overdamped, critically damped, and underdamped dynamical regimes in grokking, suggesting that the phenomenon possesses a richer phase structure than previously recognized.

## 4.7 Causal Interventions on Learning Dynamics

To test whether the observed defect accumulation and geometric reorganization are merely correlational or causally involved in grokking, we perform a series of targeted intervention experiments. These experiments modify the optimization trajectory while preserving the underlying architecture and dataset, allowing us to probe necessity and sufficiency.

### 4.7.1 Gradient Subspace Suppression

We first examine whether motion along the learned execution manifold is necessary for grokking. At each training step after step 500 (post-memorization), we project the gradient onto the subspace spanned by the top principal components of the weight trajectory, with projection strength $s \in [0, 1]$:

$$g \longrightarrow g_{\parallel} + (1 - s)\, g_{\perp}, \qquad g_{\parallel} = B\, B^{\top} g, \quad g_{\perp} = g - g_{\parallel}, \tag{8}$$

where $B \in \mathbb{R}^{P \times K}$ is the PCA basis from Phase 1 training. For comparison, we also apply random low-dimensional projections of equal rank ($K = 16$).

Partial suppression along the PCA directions ($s = 0.25$–$0.75$) systematically delays grokking, while full projection ($s = 1.0$) completely prevents generalization (0/12 seeds across four operations; Figure 13). In contrast, random projections have little effect at intermediate strengths ($< 50$-step difference from baseline; Figure 14). At $s = 1.0$, both projections kill grokking, since confining the optimizer to *any* 16-dimensional subspace of $\mathbb{R}^{290\text{k}}$ is too restrictive. These results indicate that grokking requires access to specific learned directions in parameter space, rather than arbitrary low-dimensional motion.

#### 4.7.2 Directional Forcing and Defect Induction

Next, we test whether artificially inducing curvature defects is sufficient to trigger early grokking. Starting at step 500, we periodically inject additive weight updates aligned with the commutator direction (recomputed every 50 steps), with amplitudes $\alpha \in \{50, 100, 200, 500\}$ times the gradient step norm. As a control, we apply kicks of equal magnitude along random orthogonal directions.

Across all tested amplitudes, neither commutator-aligned nor random kicks accelerate grokking relative to baseline (Figure 15). All 27/27 runs generalize at statistically indistinguishable times ($\sim$3200 steps, within seed-to-seed variability). This negative result demonstrates that defect accumulation alone is insufficient to induce grokking, and that escape from the metastable regime requires coordinated motion along learned directions.

#### 4.7.3 Replication Across Operations

We repeat the projection experiments across all four grokking tasks: modular addition, subtraction, multiplication, and quadratic addition ($a^2 + b^2 \bmod 97$). The dose–response relationship between projection strength and grokking delay is consistent across all operations (Figure 13), with complete suppression at full strength (0/12 seeds grok at $s = 1.0$). At $s = 0.75$, grokking is delayed by 600–800 steps (20–25% above baseline) across all four operations. This universality suggests that the causal role of execution-manifold directions is not task-specific, but reflects a common geometric mechanism underlying algorithmic generalization.

#### 4.7.4 Summary of Interventions

Taken together, these experiments establish a necessary–but–not–sufficient relationship between orthogonal gradient flow and grokking. Constraining motion along the execution manifold prevents generalization with a smooth dose–response curve that is specific to the PCA directions and replicates across operations; artificially increasing defect through directional forcing has no effect. This asymmetry is consistent with the commutator defect serving as a *signature* of the curvature barrier between memorization and generalization solutions—a structured reorganization of the optimization trajectory—rather than a directly manipulable cause of the phase transition. Together, these interventions rule out purely correlational explanations of our earlier findings and establish a directional causal relationship between execution-manifold geometry and generalization.

## 5 Discussion and Theoretical Connections

In this work, we investigated grokking through the lens of optimization geometry, focusing on curvature accumulation, trajectory–curvature alignment, and the emergence of low-dimensional execution manifolds. Our results suggest that grokking corresponds to a dynamical transition in parameter space, characterized by escape from a metastable, high-curvature regime into a flat, structured solution manifold. This perspective provides a unifying framework connecting grokking to several broader themes in learning theory and neural network optimization.

### 5.1 Grokking as Metastable Escape in Curved Landscapes

*Thesis: grokking is best understood as escape from a metastable regime, not merely delayed learning.*

Across tasks and hyperparameter regimes, we observe that memorization confines training trajectories to regions of high curvature anisotropy, with commutator defects gradually increasing until a critical level triggers the generalization transition. This is reminiscent of metastable escape

in stochastic dynamical systems, where the defect magnitude functions as accumulated geometric tension and the learning rate controls damping. Lower learning rates produce overdamped trajectories requiring substantial defect buildup ($\sim$30k steps at $\eta = 10^{-4}$), while higher learning rates facilitate rapid transitions ($\sim$1k steps at $\eta = 10^{-2}$; Section 4.6).

### 5.2 Implicit Regularization and Low-Dimensional Structure

*Thesis: implicit regularization in grokking operates through the emergence of task-specific geometric structure, not merely norm or margin control.*

Following grokking, trajectories collapse onto low-dimensional execution manifolds (PC1 explains 70–94% of variance; Section 4.1), accompanied by reduced curvature anisotropy along the trajectory. The intervention experiments (Section 4.7) demonstrate that motion along these specific learned directions is necessary for grokking, while generic low-dimensional constraints are insufficient. This suggests that implicit regularization in this setting operates through the progressive emergence of geometrically privileged subspaces.

### 5.3 Scaling Behavior and Phase Diagrams

*Thesis: grokking exhibits a phase diagram with power-law scaling, paralleling phenomena in larger-scale systems.*

Our learning-rate sweeps reveal distinct overdamped, critically damped, and underdamped regimes (Section 4.6), with grok time scaling approximately as $t_{\text{grok}} \propto \eta^{-1}$ and the curvature defect serving as an order parameter (sign test $p = 2^{-12}$; Section 4.4). While our experiments operate in a small-model regime ($\sim$290k parameters), these scaling relationships parallel phenomena reported in large-scale language models, suggesting that grokking may represent a microscopic instance of optimization-driven phase transitions.

### 5.4 Robustness, Flat Minima, and Quantization

Recent empirical work has demonstrated that trained neural networks can tolerate substantial parameter quantization and compression with limited performance degradation. Our results provide a geometric perspective on this robustness in the grokking regime.

Post-grokking solutions are characterized by reduced curvature anisotropy along the optimization trajectory: the commutator defect, while still nonzero, is concentrated orthogonally to the directions the optimizer traverses, and the effective integrability $\rho \approx 1.000$ (within numerical precision) indicates that the execution manifold is effectively flat at the measured resolution. Such regions, where curvature is confined to directions not visited by the optimizer, naturally support greater robustness to perturbations along the learned subspace. In contrast, pre-grokking solutions reside in highly anisotropic, high-defect regions and are correspondingly more sensitive to perturbation.

This suggests that robustness to compression may be partly understood as a consequence of geometric reorganization during training, rather than solely as a byproduct of architectural or regularization choices.

### 5.5 Connection to Mechanistic Interpretability

*Thesis: the geometric transition during grokking corresponds to the stabilization of interpretable circuits.*

The formation of execution manifolds corresponds to the concentration of computation into low-dimensional subspaces, consistent with the emergence of interpretable circuits documented in

prior work. The collapse of curvature anisotropy during grokking indicates that these circuits become geometrically stabilized. From this viewpoint, grokking marks the transition from distributed representations to structured, circuit-like organization—providing a potential bridge between optimization geometry and mechanistic interpretability.

### 5.6 Limitations and Open Problems

Our experiments are limited to relatively small Transformer models (2–3 layers, ∼290k parameters) and synthetic algorithmic tasks (modular arithmetic mod 97). While these settings permit fine-grained geometric analysis, it remains unclear to what extent the observed phenomena—rank-1 manifolds, integrable execution subspaces, predictive defect spikes—generalize to large-scale language models and real-world datasets.

In addition, several of our diagnostic measures, including commutator defects (4 forward-backward passes per sample) and trajectory–curvature alignment, are computationally expensive and difficult to scale. Developing efficient approximations and proxies for these geometric diagnostics remains an important direction for future work.

Finally, a complete theoretical characterization of the observed phase transitions remains open. Deriving analytical models that capture defect accumulation, manifold formation, and damping-controlled dynamics represents a promising avenue for future research.

### 5.7 Outlook

Taken together, our results suggest that grokking reflects a geometric reorganization of the optimization landscape, governed by curvature, damping, and emergent low-dimensional structure. By integrating dynamical, geometric, and causal analyses, this work provides a foundation for understanding delayed generalization as a phase transition in learning dynamics.

The commutator defect, in particular, offers a practical early warning signal: monitoring it during training can detect the approaching phase transition hundreds to tens of thousands of steps in advance (depending on the learning rate), suggesting substantial computational savings in settings where evaluation is expensive.

We hope that this perspective will inform future studies of optimization, scaling, robustness, and interpretability in neural networks.

## 6 Related Work

**Grokking.** Power et al. [2022] first observed delayed generalization in modular arithmetic. Nanda et al. [2023] identified "grokking circuits" (Fourier-basis representations) in 1-layer models. Liu et al. [2022] showed that grokking occurs broadly when weight decay or weight norm is controlled. Zhong et al. [2024] described clock and pizza representations in modular addition. Thilak et al. [2022] connected grokking to slingshot dynamics in adaptive optimizers. Lyu et al. [2024] characterized the role of weight decay in separating memorization from generalization phases. More recently, Merrill et al. [2023] framed grokking as competition between sparse and dense subnetworks, Varma et al. [2023] explained it through circuit efficiency, Davies et al. [2023] connected grokking to double descent, and Kumar et al. [2024] characterized the lazy-to-rich training transition. Our work complements these representational and optimization-theoretic perspectives with a geometric and causal analysis.

**Loss landscape geometry and scaling.** The study of loss landscape geometry in neural networks has a rich history [Li et al., 2018b, Draxler et al., 2018, Garipov et al., 2018]. Fort and Jastrzebski [2019] studied the curvature of the loss landscape during training. Our commutator defect is related to the Lie bracket of gradient vector fields and measures non-commutativity of the optimization flow; similar ideas appear in the study of natural gradient methods [Amari, 1998] and the Fisher information geometry of neural networks. The power-law scaling we observe in grok time vs. learning rate resonates with the broader scaling laws literature [Kaplan et al., 2020], though our analysis operates at a much smaller scale.

**Intrinsic dimensionality.** Li et al. [2018a] showed that neural network optimization occurs in a low-dimensional subspace. Xu [2026] demonstrated that attention weight trajectories during grokking in modular arithmetic lie on a low-dimensional execution manifold, with PC1 capturing the majority of variance. A corrected version of that work includes random-subspace baseline controls showing that the execution basis captures 2–10× more commutator energy than a random subspace of equal dimension. The present work extends those findings by providing evidence for effective integrability of the manifold, adding analogous random baseline controls, demonstrating that curvature dynamics predict the generalization transition, and testing the causal role of orthogonal gradient flow through intervention experiments.

# 7 Conclusion

We have shown that the weight-space trajectory during grokking lies on a rank-1, effectively integrable submanifold of parameter space, and that the principal curvature directions are predominantly orthogonal to this manifold, with curvature explosion serving as a leading indicator of the generalization transition. These findings hold across six modular arithmetic operations, three random seeds, two weight-decay settings, a 100× learning rate sweep, and two qualitatively different hyperparameter regimes (200× range in training timescale). Causal intervention experiments further establish that orthogonal gradient flow is necessary but not sufficient for grokking, with a monotonic dose–response curve that replicates across four operations and is specific to the PCA-identified directions. The geometric picture—effectively flat trajectory, orthogonal curvature, predictive spike, and asymmetric causal role—provides a new lens for understanding the grokking phenomenon and suggests that monitoring gradient non-commutativity may be useful as an early warning signal for delayed generalization in neural networks.

**Reproducibility.** All code and figures are available at `https://github.com/skydancerosel/grokking-integrability`. Total compute for full reproduction is approximately 9 hours on a single Apple M-series machine.

# References


Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

Xander Davies, Lauro Langosco, and David Krueger. Unifying grokking and double descent. *arXiv preprint arXiv:2303.06173*, 2023.

Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred A Hamprecht. Essentially no barriers in neural network energy landscape. In *International Conference on Machine Learning*, pages 1309–1318, 2018.

Stanislav Fort and Stanislaw Jastrzebski. Large scale structure of neural network loss landscapes. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry Vetrov, and Andrew Gordon Wilson. Loss surfaces, mode connectivity, and fast ensembling of DNNs. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Tanishq Kumar, Blake Bordelon, Samuel J Gershman, and Cengiz Pehlevan. Grokking as the transition from lazy to rich training dynamics. *arXiv preprint arXiv:2310.06110*, 2024.

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. In *International Conference on Learning Representations*, 2018a.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In *Advances in Neural Information Processing Systems*, volume 31, 2018b.

Ziming Liu, Ouail Kitouni, Niklas Nolte, Eric J Michaud, Max Tegmark, and Mike Williams. Omnigrok: Grokking beyond algorithmic data. *arXiv preprint arXiv:2210.01117*, 2022.

Kaifeng Lyu, Jikai Jin, Zhiyuan Li, Simon S Du, Jason D Lee, and Wei Hu. Dichotomy of early and late phase implicit biases can provably induce grokking. *arXiv preprint arXiv:2311.18817*, 2024.

William Merrill, Nikolaos Tsilivis, and Aman Shukla. A tale of two circuits: Grokking as competition of sparse and dense subnetworks. *arXiv preprint arXiv:2303.11873*, 2023.

Neel Nanda, Lawrence Chan, Tom Liberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*, 2023.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. In *ICLR 2022 Workshop on MATH-AI*, 2022. URL `https://arxiv.org/abs/2201.02177`.

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, János Kramár, and Neel Nanda. Explaining grokking through circuit efficiency. *arXiv preprint arXiv:2309.02390*, 2023.

Yongzhong Xu. Low-dimensional execution manifolds in transformer learning dynamics: Evidence from modular arithmetic tasks. *arXiv preprint arXiv:2602.10496*, 2026. URL `https://arxiv.org/abs/2602.10496`.

Ziqian Zhong, Ziming Liu, Max Tegmark, and Jacob Andreas. The clock and the pizza: Two stories in mechanistic explanation of neural networks. *Advances in Neural Information Processing Systems*, 36, 2024.

# A Additional Figures

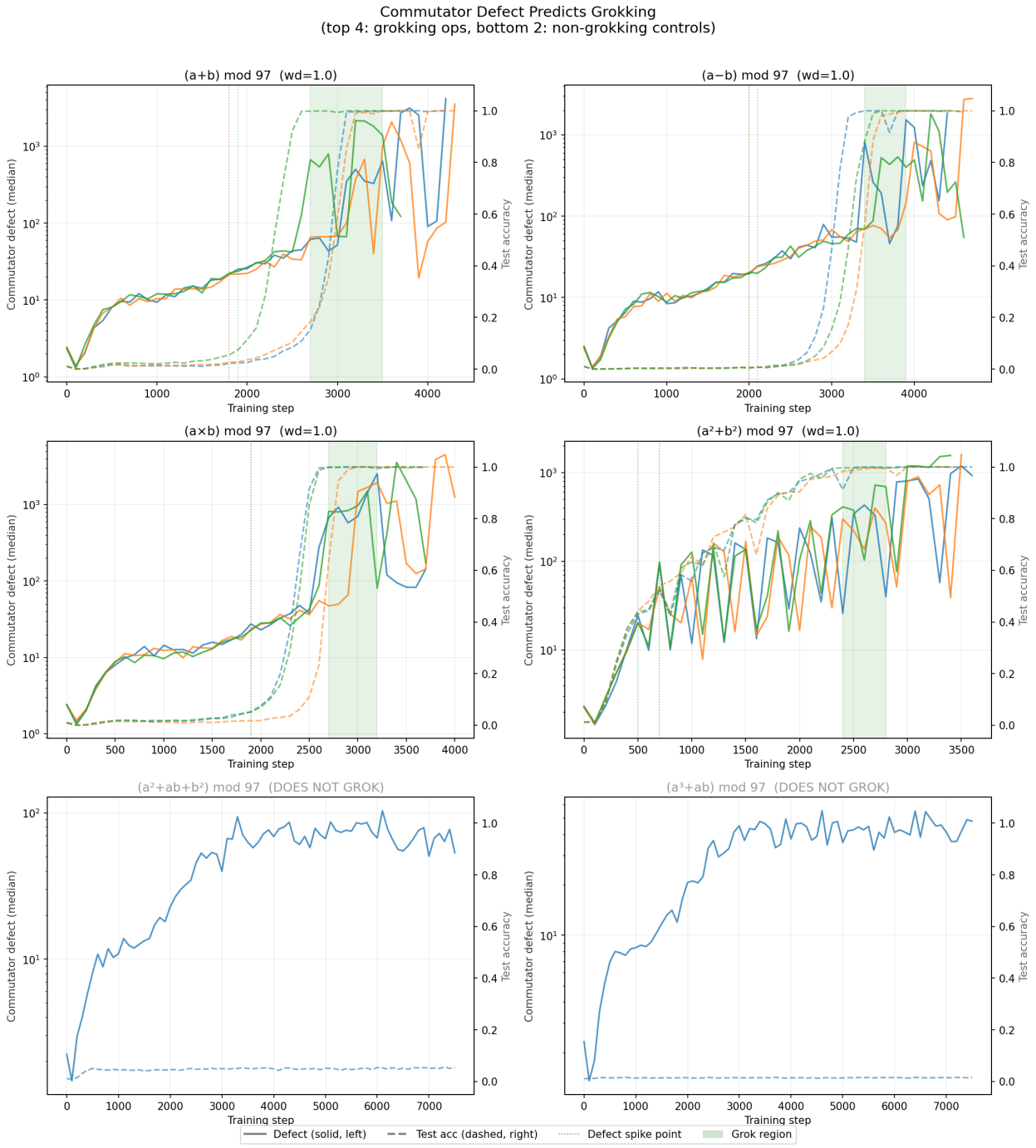
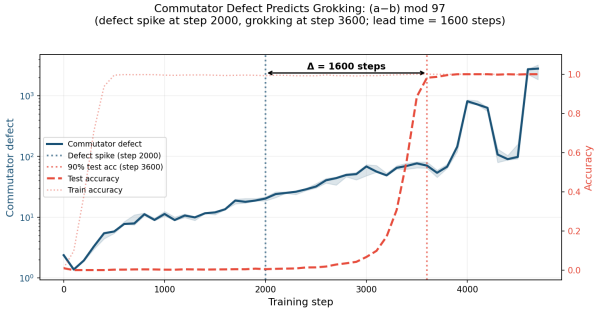
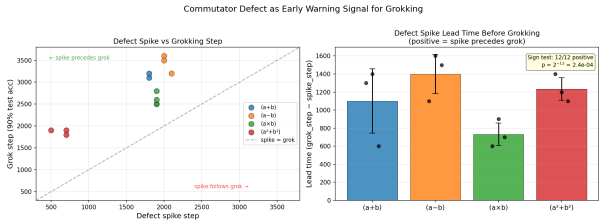



Figure 6: Commutator defect predicts grokking. Top four panels: grokking operations (3 seeds each), showing defect spike (solid) preceding test accuracy rise (dashed). Bottom two panels: non-grokking controls show moderate defect but no generalization. Dotted vertical lines mark spike detection; green regions mark grokking.

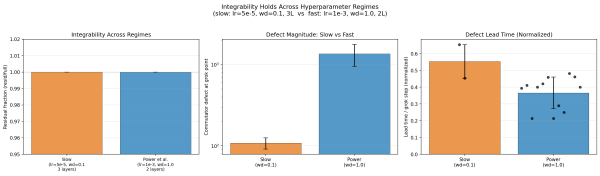



(a) Hero example: $(a - b)$ mod 97, seed 137. Defect spike at step 2000, grokking at step 3600 (lead = 1600 steps).

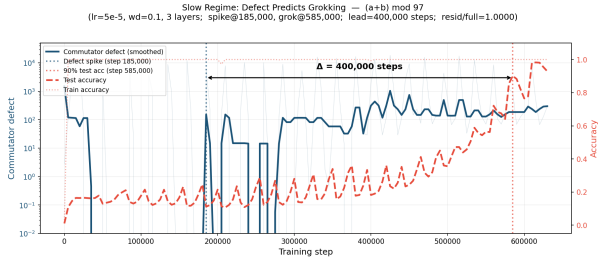



(b) Lead time quantification. Left: spike step vs. grok step (all points above diagonal). Right: lead time by operation (sign test $p < 0.001$).

Figure 7: Defect spike as early warning signal for grokking.




(a) Regime comparison: integrability, defect, and normalized lead time are consistent across regimes.




(b) Slow regime hero: defect spike at ∼185k, grokking at ∼585k (lead ≈ 400k steps).

Figure 8: Regime invariance: qualitative findings replicate in the slow regime (200× longer training), though alignment dynamics differ quantitatively.

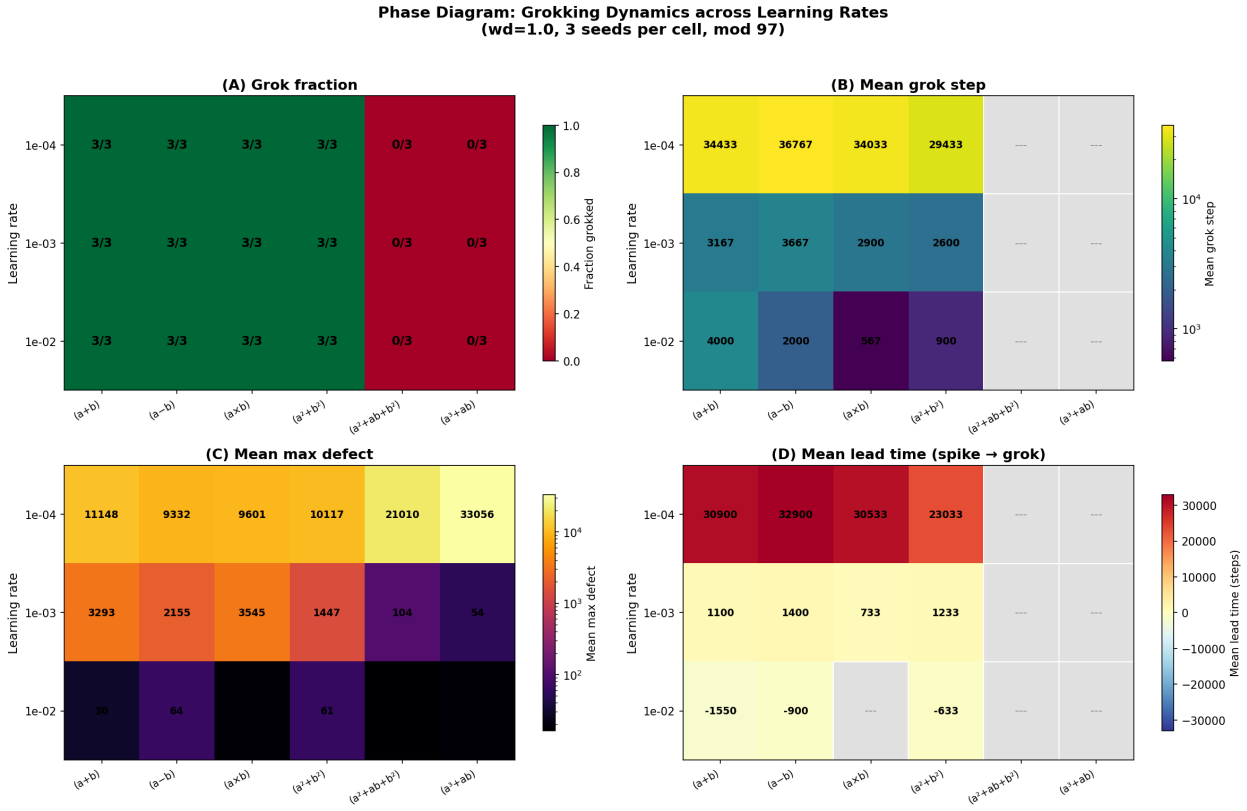



Figure 9: Phase diagram of grokking dynamics across learning rates ($\eta \in \{10^{-4}, 10^{-3}, 10^{-2}\}$, $\lambda = 1.0$, 3 seeds per cell). (A) Grok fraction: the phase boundary between grokking and non-grokking operations is invariant to learning rate. (B) Mean grok step (log scale): grokking speed scales with $\eta$. (C) Mean max defect (log scale): curvature explosion is largest at low $\eta$. (D) Mean lead time (spike step − grok step): the defect spike is most predictive at low $\eta$.

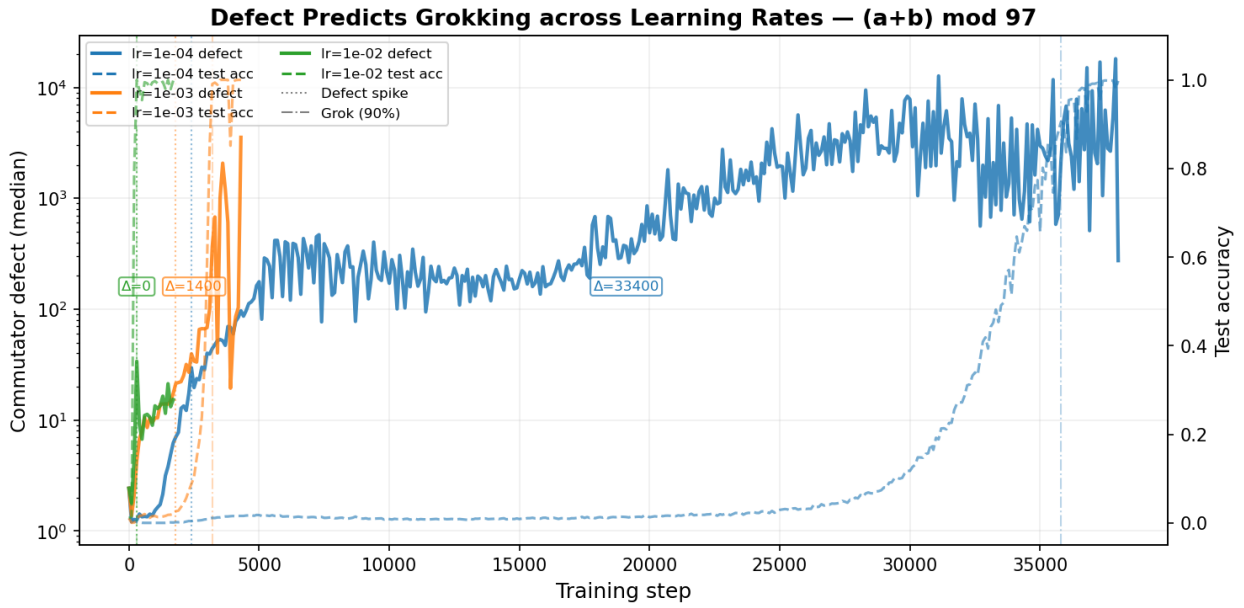



Figure 10: Defect and test accuracy trajectories for addition across three learning rates. At $\eta = 10^{-4}$ (blue), the defect spike precedes grokking by ∼33k steps; at $\eta = 10^{-3}$ (orange), by ∼1.4k steps; at $\eta = 10^{-2}$ (green), grokking is nearly instantaneous.

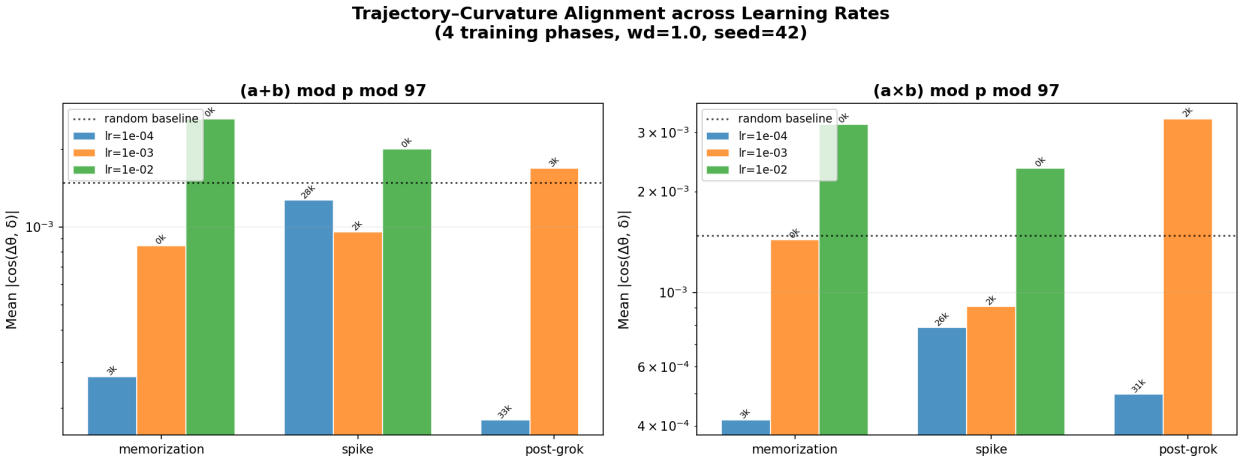



Figure 11: Trajectory–curvature alignment at three training phases across learning rates. At $\eta = 10^{-4}$ (blue), alignment stays below the random baseline (dotted), consistent with overdamped dynamics. At $\eta = 10^{-2}$ (green), alignment exceeds the baseline, consistent with underdamped exploration. $\eta = 10^{-3}$ (orange) shows intermediate behavior. Both operations exhibit the same pattern.

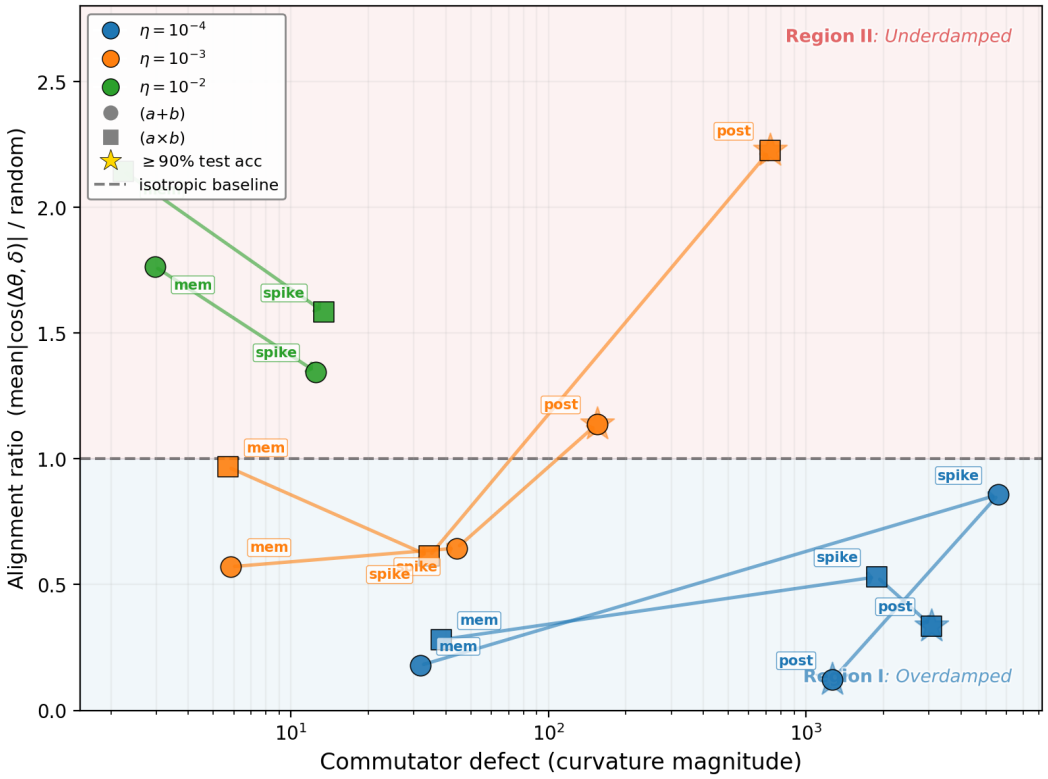



Figure 12: Phase portrait of grokking dynamics in curvature–trajectory space. We plot the trajectory–curvature alignment ratio (mean $|\cos(\Delta\theta, \delta)|$ normalized by random baseline) against the commutator defect magnitude for three learning rates ($\eta = 10^{-4}, 10^{-3}, 10^{-2}$). Each polyline traces the evolution from memorization (MEM) through the defect spike (SPIKE) to the post-grokking regime (POST), shown for addition (circles) and multiplication (squares); arrows indicate the direction of training. The horizontal dashed line indicates isotropic alignment. Training at high learning rates remains in an underdamped regime characterized by strong alignment and low defect, while low learning rates produce overdamped dynamics with large defect accumulation and weak alignment. Intermediate learning rates interpolate between these regimes. Stars mark checkpoints where test accuracy exceeds 90%. Grokking corresponds to escape from a metastable region of high curvature defect and reduced mobility, with regime-dependent relaxation dynamics.

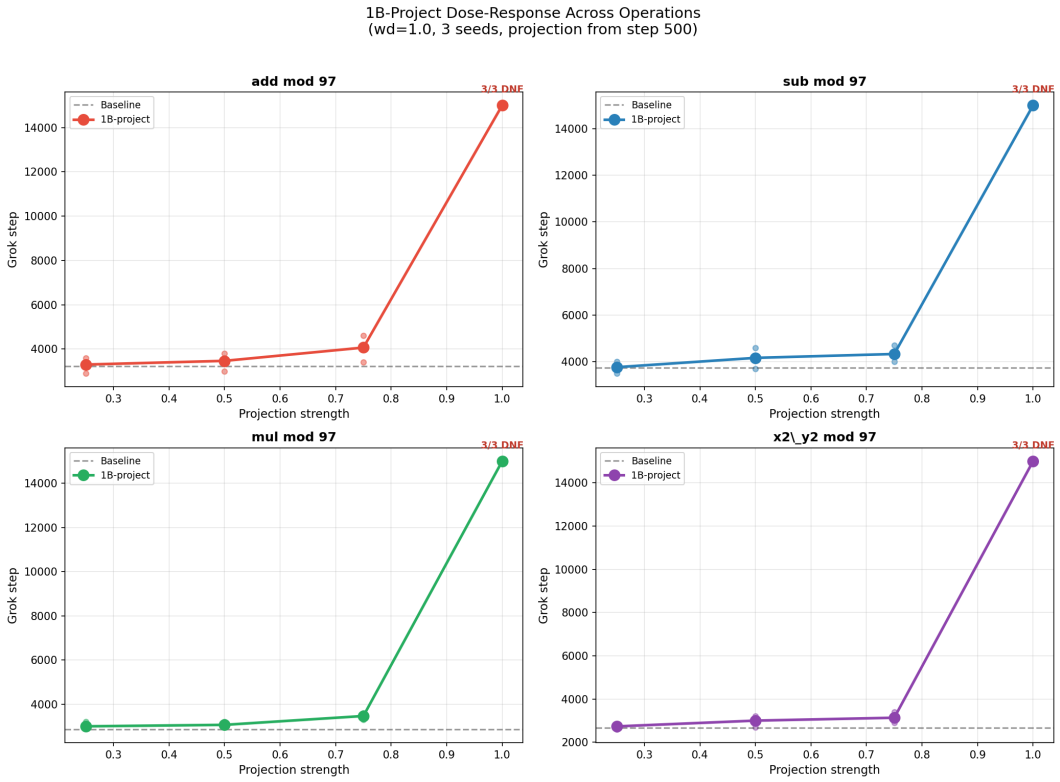



Figure 13: Dose–response curve for gradient projection across all four grokking operations. Each panel shows mean grok step (3 seeds) vs. suppression strength $s$. At $s = 1.0$, grokking fails universally (0/12 seeds). Dashed line: baseline (no intervention). The monotonic delay and complete suppression at full strength replicate across all operations.

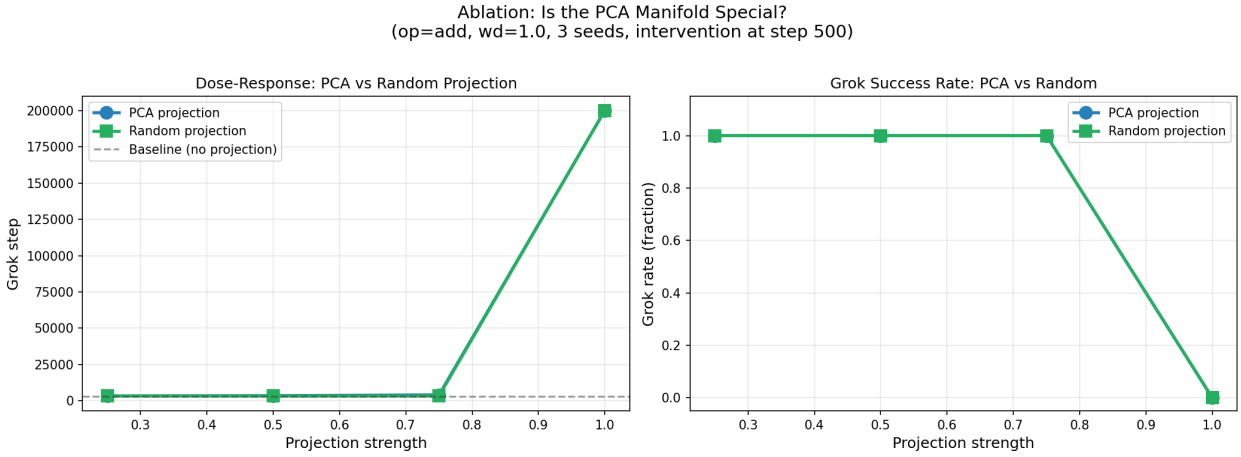



Figure 14: PCA-specific suppression control. Left: grok step vs. suppression strength for PCA projection (blue) and random projection (green); dashed line is baseline. Right: grok success rate. At intermediate strengths ($s = 0.25$–$0.75$), PCA projection monotonically delays grokking while random projection has no effect. At $s = 1.0$, both kill grokking (any 16-dim constraint is too restrictive). The dose–response separation confirms the geometric specificity of the PCA manifold.

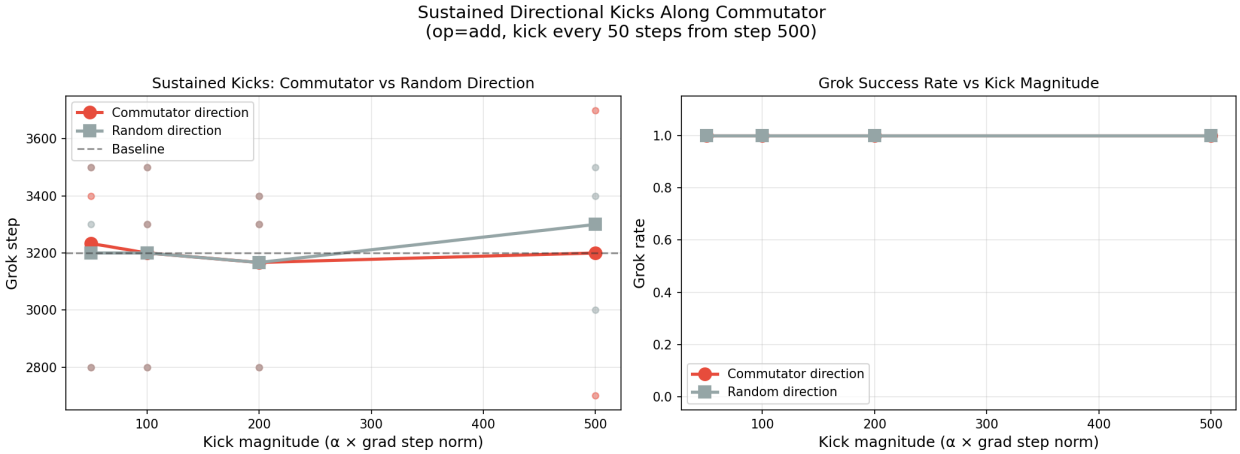



Figure 15: Sustained directional kicks along the commutator (red) vs. random orthogonal (gray) directions, with kick magnitudes up to 500× the gradient step norm applied every 50 steps. Left: mean grok step (3 seeds); right: grok success rate. Neither direction accelerates grokking beyond baseline variability (dashed line), confirming that the orthogonal defect is not sufficient to induce the phase transition.

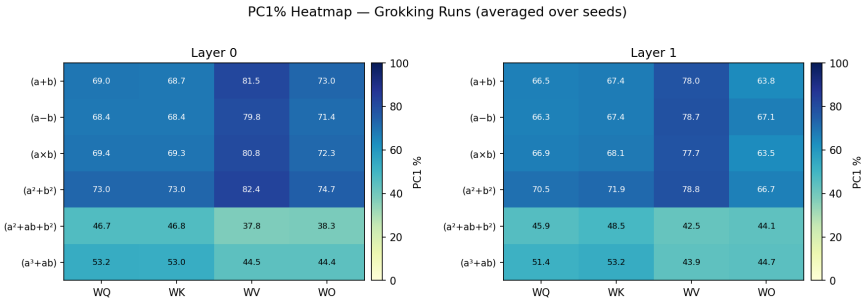



Figure 16: PC1% heatmap by operation and weight matrix (last layer, wd=1.0). All weight matrices show high PC1% for grokking operations.

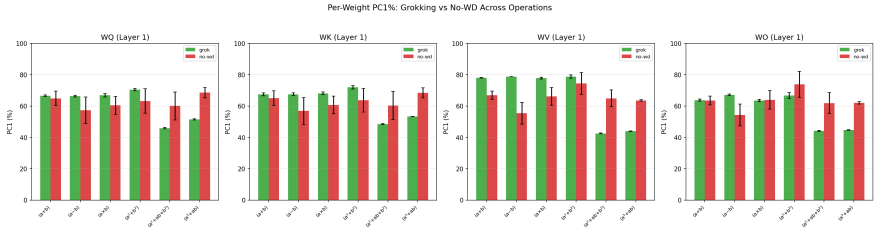
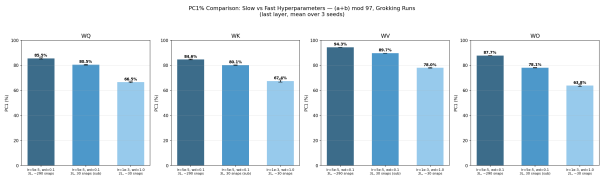
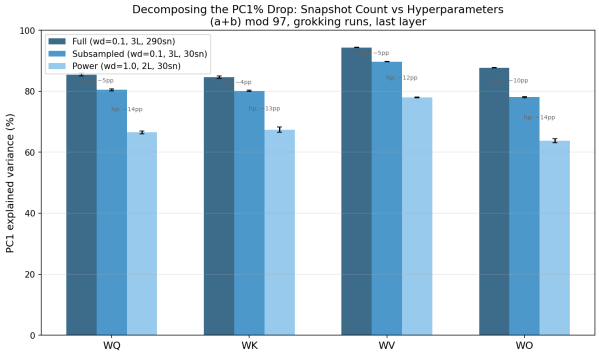



Figure 17: Per-weight-matrix PC1% comparison across operations (grok vs. no-wd).

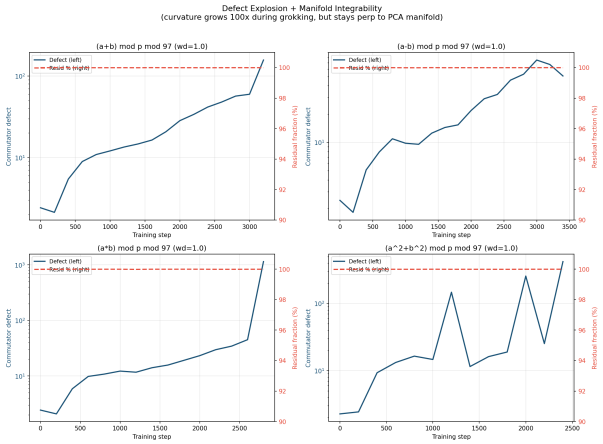



(a) Slow vs. fast regime PC1%. The slow regime shows lower PC1%, but still well above the null model.

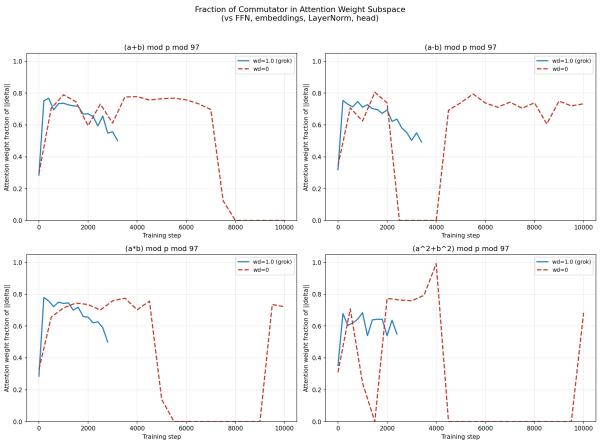



(b) Decomposition of PC1% drop between regimes: which hyperparameter drives the difference.

Figure 18: Regime comparison for PCA concentration.

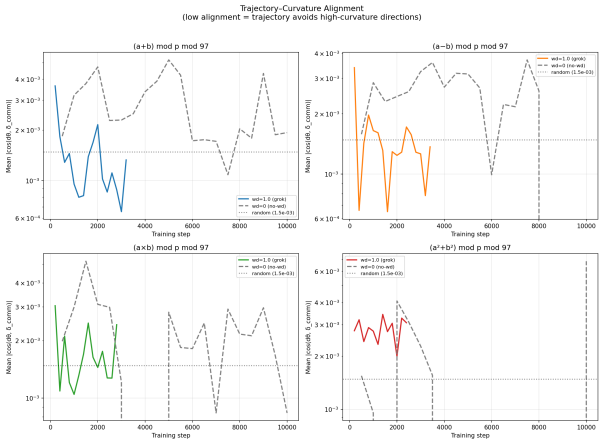



(a) Combined view: defect magnitude and integrability over training.

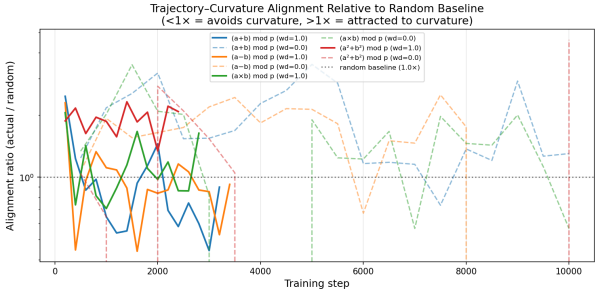



(b) Attention weight fraction of commutator defect.

Figure 19: Commutator analysis details.




(a) Trajectory-curvature alignment over training.




(b) Alignment ratio vs. random baseline: the trajectory does not prefer curvature directions.

Figure 20: Converse analysis: the weight trajectory avoids high-curvature directions.

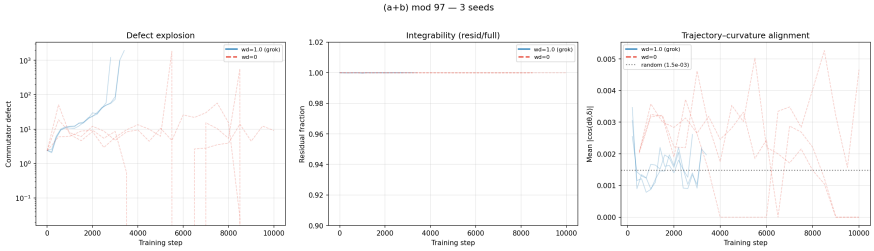



Figure 21: Temporal traces for $(a + b) \bmod 97$ with 3 seed overlays, showing consistency of the integrability and defect patterns across seeds.