# Grokking Collapses the Algorithm, Not the Function

## Irrep-Energy Underdetermination in Modular Addition

**Kamil Dixon**

June 2026 Third Rail

# Introduction

A small neural network trained on modular addition with sufficient weight decay eventually transitions from memorizing its training set to generalizing on its held-out set, sometimes long after training loss has saturated. This phenomenon, called grokking (Power et al. 2022), has become a controlled testbed for studying how networks acquire generalizable algorithms. Subsequent mechanistic work (Nanda et al. 2023; Gromov 2023) showed that grokked networks discover a Fourier-based solution: they encode inputs at a small set of frequencies, perform multiplication in their middle layer, and decode the result by reading out the angle. The grokking phase transition coincides with the emergence of this Fourier circuit.

Most existing analysis treats grokking as the network finding *the* generalizing solution. The Fourier circuit is described as universal; the post-grokking endpoint is described as a singular point in algorithm space. This paper asks whether that singularity is real.

We trained 217 grokked modular-addition networks across 10 random seeds and a sweep of weight-decay strengths, learning rates, and decay schedules, and measured the geometry of the resulting population. The grokking phase transition collapses test accuracy: every model reaches accuracy $\geq 0.95$ on the held-out set. It does not collapse the network. The 217 grokked models, each implementing the same modular-addition algorithm with the same Fourier scaffold, occupy a logit-function family of low effective dimension (PR $\approx 6$ in the original-batch analysis of §6), while hidden activations and parameters spread over substantially more directions (PR $\approx 74$ and PR $\approx 95$–$110$ respectively, same analysis). Models share the algorithm; they do not collapse to a single logit-function realization.

The question this paper answers is: *what coordinate system parameterizes the post-grokking function family?* If the family is low-dimensional, there should be compact coordinates that describe where each model sits within it. We sought those coordinates, and we identify a compact, interpretable coordinate system that recovers much of their held-out geometry.

We introduce three terms that organize the paper. The **scaffold** is the shared Fourier circuit class to which grokked networks converge — the cross-and-diagonal signature in hidden activations that the analyzed grokked population shares, irrespective of seed or schedule. The **dialect** is the residual real-valued logit-function variation within the scaffold: same modular-addition behavior above the grokking threshold, different confidence geometry. Each model speaks the same Fourier language but writes different logit values into it. The **dialect subspace** is the coordinate system that indexes those values — a compact 12-to-24-dimensional region of readout space, revealed by held-out cross-seed evaluation, that captures the transferable geometry of the function family.

### Contributions

1. **The post-grokking function family is structured.** Across 217 grokked models, hidden-activation Fourier power concentrates on the same algorithmic signature (cross lines and

diagonals) with the same dominant frequencies; given grokking, the Fourier scaffold is shared. Yet the logit-function family is structured rather than degenerate: in the original-batch PR analysis of §6, the family occupies an effective subspace of $\approx 6$ directions, distinguishing post-grokking from "all grokked models implement the same function" and from "grokked models occupy independent isolated solutions."

2. **Cross-seed coordinates are not in the hidden-layer Fourier amplitudes.** A 24-cell selected Fourier feature set captures population geometry strongly within a pooled sample (Mantel 0.65) but fails leave-one-seed-out evaluation (gap 0.10, $z = 0.71$). A broader full top-power Fourier variant similarly fails cross-seed (gap 0.09, $z = 0.77$). The hidden-layer Fourier description identifies the algorithm but does not parameterize the dialect across initialization basins.

3. **The dialect subspace is recoverable from, and partially controllable through, the readout layer, in a structurally interpretable basis.** Held-out cross-seed decoding from the readout matrix reaches gap 0.41 at $z = 2.0$. Dimensionality-matched controls (PCA-readout-24 matches the full 6,016-dimensional readout) rule out feature-count advantage. A 47-dimensional gauge-invariant aggregate — the total Fourier power per irrep of $\mathbb{Z}/47\mathbb{Z}$ summed across hidden units — recovers held-out function geometry at least as well as the full readout. We call this aggregate the *irrep-energy budget* and identify it as the coordinate system parameterizing the dialect subspace.

4. **Surgical intervention on the irrep budget moves function in predicted directions.** For 20 grokked models, three frequencies (two dominant, one low-power control), and five scaling factors, we modify the readout's irrep coefficients and measure the resulting logit-function shift. At dominant frequencies, predicted and observed shifts agree on the non-leading function-PC subspace (mean cosine $+0.62$ at $k = 4$, $+0.32$ at $k = 11$, with median values near $+0.99$); at the control frequency, no systematic agreement is observed (mean cosine $-0.23$, no interventions reach the strong-agreement threshold). Grokking is preserved across all interventions. The irrep budget is partially controllable, not just decoded.

5. **The dialect subspace is causally asymmetric under parameter perturbation.** Across 217 grokked models, perturbations along the dialect direction — the readout-Fourier intervention from claim 4, normalized — preserve test accuracy across four orders of magnitude in parameter L2 norm. Random parameter directions, fc2-locality-matched random directions, and scaffold-sensitive directions all break the function at perturbation norms in the 8–12 range. Hidden-unit permutation, an exact functional invariance, preserves test accuracy to numerical precision (sanity check). The dialect direction is sharply separated from generic parameter perturbation, and the difference cannot be attributed to fc2-locality alone — fc2-random tracks full-parameter random closely. The function family has a function-preserving direction whose extent is measurable.

6. **Methodological recommendation.** The within-population-to-cross-seed gap we observe in claim 2 is large — Mantel 0.65 collapses to gap 0.10 on the same feature set. Representational-mechanism claims that pass within-population tests need not pass held-out cross-seed tests. We recommend leave-one-seed-out evaluation with shuffled-target nulls, and falsification via control-feature interventions, as standard supplements to in-sample correlations in mechanistic interpretability work.

### What this paper does not claim

We restrict our claims in three ways. First, we claim geometry recovery rather than coordinate prediction; per-PC $R^2$ is weak across folds. Second, we claim measurable causal asymmetry in the dialect subspace, not complete causal identification of its internal coordinate structure; intervention magnitudes are damped, suggesting the budget is a coupled property of the full network. Third, we claim a mechanism within this system, not universality across tasks; we test one task, one architecture, and one prime. §11 returns to these boundaries in detail.

### Paper organization

§3 specifies the experimental setup and the cross-validation protocols used throughout. §4 establishes local mode connectivity between same-seed grokked endpoints, addressing whether population dispersion should be interpreted as isolated basins. §5 shows the shared Fourier scaffold across our grokked population. §6 reports the dimension ladder distinguishing accuracy, logit-function, hidden-activation, and parameter spaces. §7 shows that hidden-Fourier features explain seed-local geometry but fail cross-seed transfer. §8 establishes that the cross-seed signal is recoverable from the readout layer and is dimensionality-compact. §9 names the readout coordinate as the irrep-energy budget, characterizes its symmetry and effective dimension, reports the intervention experiment that supports its causal role (§9.4), and shows the dialect direction is causally asymmetric under parameter perturbation (§9.5). §10 discusses the synthesis, why it matters methodologically, and open questions. §11 consolidates the boundaries and limitations of all claims.

## Related work

### Grokking dynamics and Fourier mechanisms

The grokking phenomenon — networks that transition from memorization to generalization long after training loss has saturated — was reported by Power et al. (2022) on small algorithmic tasks including modular arithmetic. Subsequent work has characterized the dynamics of this transition (Liu et al. 2022; Thilak et al. 2022; Žunkovič and Ilievski 2022), the role of weight decay and regularization in inducing it (Liu, Michaud, and Tegmark 2023), the early prediction of grokking from loss-landscape structure (Notsawo et al. 2023), accounts of competing memorizing and generalizing circuits that govern its timing (Varma et al. 2023), and a treatment of grokking as a first-order phase transition (Rubin, Seroussi, and Ringel 2024).

On the mechanism side, Nanda et al. (2023) showed that grokked networks on modular addition implement a Fourier-based algorithm: inputs are encoded as $x \mapsto e^{2\pi ikx/p}$ at a small set of frequencies, multiplied by trigonometric identity through the hidden layer's nonlinearity, and decoded by reading out the angle. Gromov (2023) provided analytic feature-map reductions of trained networks for modular arithmetic, and Zhong et al. (2023) analyzed alternative mechanistic explanations including the "clock" and "pizza" representations that grokked networks can produce. Together this literature describes when and how grokking happens, and what shared algorithmic scaffold grokked networks implement. We condition on this scaffold appearing.

### Universality and population-level circuit structure

Recent work has studied whether group-operation circuits recur across architectures, seeds, and groups, including representation-theoretic analyses of universality in group-operation networks

(Chughtai, Chan, and Nanda 2023). Our paper asks a complementary question. We condition on the shared scaffold appearing and study the degrees of freedom left within that scaffold: which function-level coordinates vary across grokked models, where those coordinates are represented, and whether interventions on those coordinates produce predicted function-space shifts.

### Mode connectivity and solution geometry

The geometry of trained-network solutions has been studied through the lens of mode connectivity: Garipov et al. (2018) and Draxler et al. (2018) found that trained networks at apparent local minima are typically connected by curved low-loss paths, despite the linear interpolation between them crossing barriers. Frankle et al. (2020) studied linear mode connectivity as a probe of training dynamics and initialization-dependent solution structure; subsequent work has connected this to the geometry of generalization strategies (Juneja et al. 2023). Our §4 shows the same pattern in our grokked endpoints at the same-seed scale, and we use that result to caution against interpreting our population-level dispersion as dispersion across isolated basins.

### Mechanistic interpretability, interventions, and representation comparison

Mechanistic interpretability work typically reverse-engineers circuits inside individual trained models or fixed model classes, including visual circuits (Olah et al. 2020), transformer circuits (Elhage et al. 2021), induction heads (Olsson et al. 2022), and automated circuit discovery (Conmy et al. 2023). Cross-model representation comparison has been developed in parallel: model stitching and centered kernel alignment (Lenc and Vedaldi 2015; Bansal, Nakkiran, and Barak 2021; Kornblith et al. 2019) compare what representations encode across architectures and training conditions. This paper draws on tools from both threads: targeted intervention on a single subspace, and decoder-based comparison of representations across initialization seeds.

### The distinction

Prior work explains the scaffold or compares trained representations. This paper studies the residual degrees of freedom left after the scaffold appears, identifies a transferable readout coordinate — the irrep-energy budget — for that variation, and tests that coordinate with controlled interventions.

## Setup

### Task and data

We study modular addition over $\mathbb{Z}/p\mathbb{Z}$ with $p = 47$. The input space is $\{0, \ldots, 46\}^2$ ($P^2 = 2{,}209$ pairs); the target is $y = (a+b) \bmod p$. We use a fixed train/test split of 40%/60%, randomly drawn once with a global seed, yielding $|\text{train}| = 883$ and $|\text{test}| = 1{,}326$. The same split is used for every model in the population so that the held-out test set is identical across all 217 models we analyze.

### Architecture

Each model is a two-layer MLP with one hidden ReLU layer:

$$\hat{y} = \text{fc}_2(\text{ReLU}(\text{fc}_1(\text{embed}(a) \,\|\, \text{embed}(b)))),$$

with embedding dimension 32 (per-input), hidden width $h = 128$, and output dimension $p = 47$. Concatenated embedding vectors enter $\text{fc}_1$ as 64-dimensional inputs. Total trainable parameters: 15,887.

### Optimizer and weight-decay schedules

Models are trained with AdamW, full-batch, using cross-entropy loss, for schedule-dependent step budgets of up to 20,000 steps. Original-batch constant-schedule runs use 8,000 steps; original-batch cyclic-schedule runs use 2,000 plus three schedule periods, capped at 20,000; extension-batch runs use 6,000 steps. Appendix A gives the per-batch schedule and step-budget details. The grokking threshold of 0.95 is reached within the relevant budget for the 217 models we analyze. Two schedule families are used:

- **Constant** schedule: weight decay $\lambda \in \{0.3, 0.5, 1.0, 1.5, 2.0, 3.0\}$ held constant throughout training, learning rate $\eta \in \{0.005, 0.01, 0.02\}$.
- **Cyclic** schedule: weight decay oscillates sinusoidally between $\lambda_{\min}$ and $\lambda_{\max}$ with a fixed period; learning rate $\eta = 0.01$ throughout. Original-batch cyclic configurations use $\lambda \in [0.3, 3.0]$ and $\lambda \in [0.5, 2.5]$ across periods $\{1{,}500, 3{,}000, 6{,}000\}$ steps; extension-batch cyclic configurations use $\lambda \in [0.3, 2.5]$ and $\lambda \in [0.5, 2.0]$. Per-batch cyclic-configuration counts are in Appendix A.4.

These ranges were chosen to span the regime in which grokking occurs reliably for this architecture.

### Population construction and grokking filter

We trained models across 10 random seeds — five original ($\{7, 13, 21, 42, 99\}$) and five extension seeds ($\{101, 137, 211, 313, 401\}$) — paired with the schedule grid above. Across both batches, 217 of 232 attempted runs reached our grokking threshold of test accuracy $\geq 0.95$ within their respective step budgets; the 15 non-grokked runs cluster in the low-$\lambda$, low-$\eta$ corner of the schedule grid. The 217 grokked models constitute our analyzed population. Unless otherwise stated, subsequent decoding evaluations and intervention experiments use this 217-model grokked population; §6 reports raw participation-ratio statistics on the original 117-model batch for the scale-heterogeneity reasons documented in Appendix A.5.[1]

### Cross-model alignment

To compare hidden representations and parameters across models without confounding by unit-permutation gauge, we align each model's hidden units to a fixed reference model: model 0, the first grokked model in the frozen population ordering. Alignment uses the activation-Hungarian procedure: for each pair (reference, target), we compute the cosine-similarity matrix between their normalized hidden activations on the full input grid, then apply the Hungarian assignment algorithm to find the permutation of target hidden units that maximizes total similarity to the reference. The same permutation is applied jointly to fc1's output units and the corresponding hidden-unit axis of fc2 when reporting parameter-space and readout-matrix quantities. Embedding vectors are not permuted because they precede the hidden layer.

### Logit-function space and PC basis

For each model we record the logit output on the held-out test split, flattened to a vector $\ell \in \mathbb{R}^{P \cdot |\text{test}|} = \mathbb{R}^{62{,}322}$. We construct a six-dimensional logit-function principal-component basis by stacking these vectors across all 217 models, mean-centering across the population, and taking the leading six right-singular vectors of the resulting matrix. The six PCs span the dominant directions

[1] Population details, schedule grids, and per-batch grokking counts are given in Appendix A.

of population variance in logit space and are the basis we use throughout the paper when discussing function-space coordinates.

For per-fold cross-seed evaluations (§§7, 8, and 9.1–9.3), the PC basis is recomputed within each training fold using only the training subset of models, then applied to the held-out seed's logits to obtain held-out coordinates.

### Held-out seed protocol

To test whether features identified within one population of grokked models transfer to a different population, we use leave-one-seed-out cross-validation: each of the 10 seeds is held out in turn, the remaining 9 seeds are used to fit features and a decoder, and the held-out seed is used as the test set. This is the protocol used in §§7, 8, and the decoding analyses in §9.1–§9.3. The intervention experiment in §9.4 is a within-population causal probe rather than a cross-seed generalization test; it therefore uses the full-population PC basis, as described in §9.4.

For each fold, we generate a shuffled-target null distribution by permuting the training-fold labels 100 times before fitting the decoder. The resulting Mantel correlations form the null against which the real-target Mantel is compared.

### Metrics

We use three quantitative measures throughout the paper:

- **Participation ratio** (PR): $\mathrm{PR}(X) = \dfrac{\left(\sum_i \mu_i\right)^2}{\sum_i \mu_i^2}$, where $\mu_i$ are the eigenvalues of a population covariance matrix. Used in §6 to summarize effective dimensionality.
- **Mantel correlation**: the Pearson correlation between two pairwise-distance matrices on the same set of items, used to test whether one feature space preserves the geometric structure of another. We report Mantel correlations between predicted-coordinate distances and true-coordinate distances on held-out test models.
- **Held-out Mantel gap and z-score**: the gap is the real-target Mantel minus the mean of the shuffled-target Mantel distribution; the z-score is the gap divided by the standard deviation of the shuffled-target distribution. We use the conventional thresholds *weak* ($\geq 0.05$), *meaningful* ($\geq 0.15$), and *strong* ($\geq 0.30$) for the Mantel gap, and report the z-score for shuffle-relative significance.

## Local mode connectivity

Before measuring the geometry of the population in §§5–9, we establish a local-curvature baseline. The question we settle here is whether two grokked models from the same initialization seed but different weight-decay configurations are linearly disconnected — separated by a high-loss barrier on the straight-line path between them — or whether low-loss curved paths connect them. The answer affects how the population-level dispersion reported in §6 should be interpreted.

For each of five seeds we selected two grokked endpoints with different weight-decay configurations. For each endpoint pair we then measured cross-entropy loss along two paths: a straight-line linear interpolation $\theta(t) = (1-t)\theta_a + t\theta_b$ for $t \in [0, 1]$, and a curved quadratic Bezier interpolation through a learned midpoint $\theta_m$ chosen to minimize the maximum loss along the path. Both paths share

endpoints; only the trajectory between them differs. Endpoint pair selection criterion, per-seed configurations, and Bezier optimization details are given in Appendix B.

The two paths produce qualitatively different loss profiles (Figure 1). Linear interpolation crosses barriers between 2.5 and 3.0 in cross-entropy, while the Bezier path stays below 0.2 barrier height for all five seeds.

- **Linear interpolation** crosses a substantial barrier. Maximum cross-entropy along the straight line ranges from 2.54 to 2.99 across the five seed pairs (mean 2.75). Test accuracy along the linear path drops below the grokking threshold over an interval of $t$ around the midpoint.
- **Curved interpolation** through the optimized midpoint reduces the barrier height to below 0.20 on every seed pair, and below 0.10 in four of five. Test accuracy along the Bezier path remains above the grokking threshold throughout.

The midpoint is not a separately trained model — it is found by gradient descent on the path's maximum-loss objective with the endpoints fixed. The fact that such a midpoint exists, and that the resulting curved path stays in the low-loss region, shows that the two grokked endpoints are connected by at least one low-loss curved path in their joint neighborhood.

This result is consistent with the broader mode-connectivity literature, in which curved low-loss paths between trained networks are routinely found despite the linear interpolation between them crossing barriers (Garipov et al. 2018; Frankle et al. 2020). We confirm here that grokked solutions on this architecture inherit that property locally.

The implication for §§6–9 is methodological. The population-level dispersion we will report in §6 — measured on the original 117-model batch for the scale-heterogeneity reasons documented in Appendix A.5 — shows logit-function PR $\approx 6$, hidden-activation PR $\approx 74$, and parameter PR $\approx 95-110$. Without this section, a reader could plausibly interpret that dispersion as evidence for *disconnected* basins, in which case the cross-seed coordinates in later sections would be coordinates of basin identity rather than coordinates within a continuous family. §4's result cautions against this reading: at least locally, between same-seed endpoints, low-loss curved paths connect grokked solutions. Subsequent dispersion measurements should not be assumed to reflect isolated points.

We do not claim that this connectivity extends globally across all 217 models. Cross-seed connectivity, in particular, is harder to establish and we do not address it here. The local result is sufficient for the methodological purpose: subsequent dispersion measurements are not a priori dispersion across discrete basins.

---

## A shared Fourier scaffold

Modular addition $a+b \pmod p$ has a well-known Fourier solution: encode $a$ and $b$ as unit vectors in $\mathbb{C}^p$ at frequency $k$ via $a \mapsto e^{2\pi i k a/p}$, multiply, and read out the angle. Prior work (Nanda et al. 2023; Gromov 2023) showed that grokked networks discover this solution by allocating Fourier mass at a small set of frequencies and implementing the trigonometric product structure in their middle layer. Here we verify that this discovery is shared at population scale: the analyzed grokked population exhibits the same cross-and-diagonal Fourier scaffold, and we use the resulting structure as the shared alphabet whose realization within initialization basins the rest of the paper investigates.

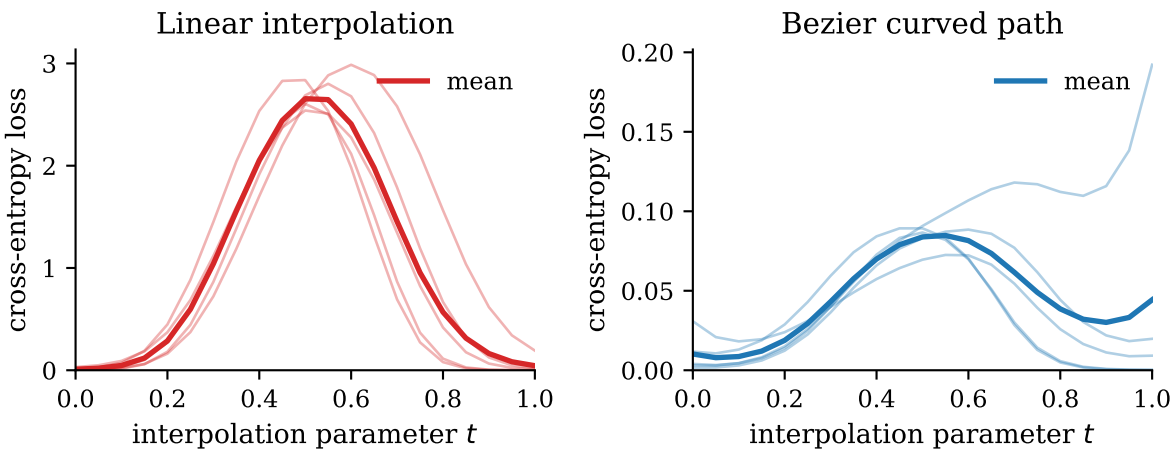


Figure 1: *Local mode connectivity between same-seed grokked endpoints.* Cross-entropy loss as a function of interpolation parameter $t$ along (left) the linear path $\theta(t) = (1-t)\theta_a + t\theta_b$, and (right) a quadratic Bezier path through an optimized midpoint $\theta_m$. Five seeds shown; mean and individual curves displayed. The linear path crosses a cross-entropy barrier between 2.5 and 3.0, while the Bezier path stays below 0.2 barrier height throughout. Endpoint pair configurations and Bezier optimization details are in Appendix B.

This reproduces the known grokking Fourier signature at population scale; the novelty is what remains variable after this scaffold appears.

For each grokked model we computed the post-ReLU hidden activations $H_{ab} \in \mathbb{R}^h$ on every input pair $(a, b) \in \{0, \dots, p-1\}^2$ ($p^2 = 2{,}209$ pairs total) and applied the two-dimensional discrete Fourier transform over the input grid:

$$\widehat{H}[k_a, k_b, j] = \sum_{a,b} H_{ab}[j]\, e^{-2\pi i (k_a\, a + k_b\, b)/p}, \qquad (k_a, k_b) \in \{0, \dots, p-1\}^2.$$

The squared magnitude $|\widehat{H}[k_a, k_b, j]|^2$ summed across hidden units $j$ gives the Fourier power that a model places on the input pattern $(k_a, k_b)$. Averaging across the 217 grokked models yields a population-level Fourier-power map.

The map concentrates on a sparse, structured pattern (Figure 2):

- **Cross lines** $k_a = 0$ and $k_b = 0$, capturing the marginal Fourier content of each input.
- **Diagonals** $k_a = k_b$ and $k_a = -k_b \pmod p$, capturing the sum and difference channels that the modular-addition algorithm requires.
- **A small set of dominant frequencies** along these signature lines, with the largest mass at conjugate pairs $k = 4 \leftrightarrow 43$ and $k = 11 \leftrightarrow 36$, followed by smaller mass at additional pairs.

Off-signature regions of the frequency plane carry much less power: the population mean drops by more than two orders of magnitude away from the cross-and-diagonal pattern. The same pattern appears consistently across individual models — the population mean is not the result of averaging over disagreement; model-level Fourier-power maps carry the same scaffold, with differences in the relative weight assigned to the dominant frequencies.

Across the 217 grokked models we tested — spanning 10 random seeds, 6 weight-decay strengths, 3 learning rates, and both constant and cyclic decay schedules — the cross-and-diagonal signature appears consistently, with dominant mass recurring at the same conjugate frequency pairs. This regularity is conditioned on grokking: models that did not reach acc $\geq 0.95$ are not in the analyzed population. Our claim is that *given grokking*, the Fourier scaffold is shared.

This section establishes the alphabet — the *scaffold* in the terminology of §1. The remainder of the paper concerns the family of functions that this alphabet can spell — the *dialect*. We show in §6 that grokked models, despite sharing the scaffold, occupy a logit-function space of effective dimension $\approx 6$ rather than collapsing to a single function, and we show in §§7–9 that the coordinate which indexes position within the dialect is not any single Fourier amplitude but a gauge-invariant aggregate of the readout layer.

---

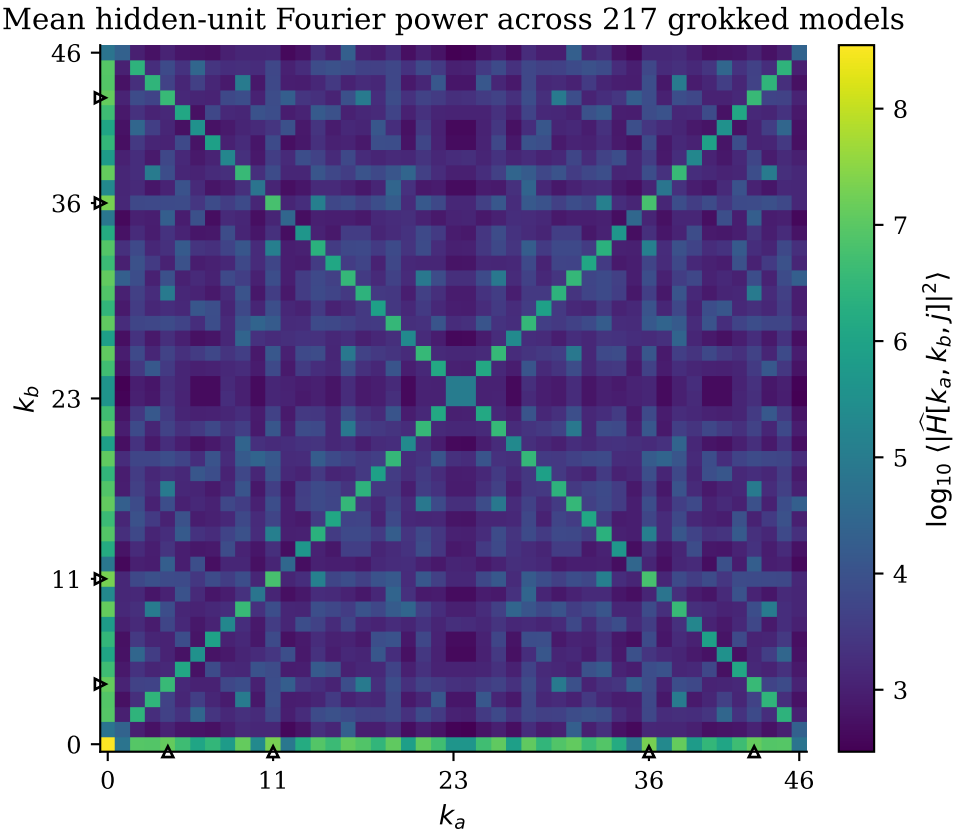


Figure 2: *The Fourier scaffold of grokked modular-addition networks.* Mean log-power $\log_{10}\langle|\widehat{H}[k_a, k_b, j]|^2\rangle_{j,\,\text{models}}$ across the input frequency grid, averaged over hidden units and over all 217 grokked models. Power concentrates on the cross lines $k_a = 0$ and $k_b = 0$ and on the diagonals $k_a = \pm k_b \pmod p$, with peaks at conjugate-pair frequencies $\{4, 11, 36, 43\}$ and smaller mass at additional pairs. Off-signature regions are more than two orders of magnitude lower. The same structure is present across individual models in the population.

## Behavior, function, and parameter dimensions decouple

§5 established that grokked models share a Fourier scaffold in their hidden activations. We now ask whether the convergence implied by "shared scaffold" extends to convergence in higher-dimensional measures of model state. To answer this we measure the effective dimensionality of four observable spaces across the 117 grokked models of the original training batch (App A.5); §§7–9 use the full 217-model population.

### Four observable spaces

For each grokked model we compute four representations:

- **Test accuracy**, a scalar in $[0, 1]$ measuring the fraction of held-out inputs on which the model predicts the modular-addition target.
- **Logit-function**, the model's logit output on the held-out test split, flattened to a vector in $\mathbb{R}^{P \cdot |\text{test}|}$ ($47 \cdot 1326 = 62{,}322$ scalars). This represents the model's functional behavior beyond the binary correct/incorrect distinction.
- **Hidden activations**, the post-ReLU hidden-layer responses on the full input grid, flattened to a vector in $\mathbb{R}^{P^2 \cdot h}$ ($2209 \cdot 128 = 282{,}752$ scalars). This represents what the network internally computes.
- **Parameters**, the flattened weights of all three layers (embedding, fc1, fc2), totaling 15,887 scalars per model. This represents the model's location in weight space.

Hidden activations and parameters are aligned across models by the activation-Hungarian matching procedure described in Appendix A; we report dimensionality on the aligned representations to remove the trivial unit-permutation gauge.

### The participation ratio

To compare effective dimensionalities across spaces of different ambient sizes, we use the participation ratio

$$\mathrm{PR}(X) = \frac{\left(\sum_i \mu_i\right)^2}{\sum_i \mu_i^2}$$

where $\mu_i$ are the eigenvalues of the cross-model covariance matrix. PR ranges from 1 (all variance on a single direction) to the ambient dimension (variance spread uniformly). It is invariant to the ambient dimension, making cross-space comparison meaningful.

### The ladder

Across the 117-model original-batch population:

@ >

p(- 6) * 0.2500 >

p(- 6) * 0.2500 >

p(- 6) * 0.2500 >

p(- 6) * 0.2500@

Space

&

Ambient dim

&

PR

&

PR / ambient

Test accuracy & 1 & $\sim 1$ & —
Logit-function & 62,322 & 6.5 & $1.0 \times 10^{-4}$
Hidden activations (aligned) & 282,752 & 74 & $2.6 \times 10^{-4}$
Parameters (aligned) & 15,887 & 95−110 & 6–7 $\times 10^{-3}$

(Test-accuracy variance is collapsed by our grokking filter at acc $\geq 0.95$; we list it as scalar-dimensional in the formal sense but the population spread is small and noise-dominated.)

These dimensions are decoupled by orders of magnitude. Logit-function variation occupies an effective subspace of $\sim 6$ directions in a 62,322-dimensional ambient space — a compression by four orders of magnitude. Hidden activation variation is broader (PR $\sim 74$), and parameter variation is broader still (PR $\sim$ 95–110). Accuracy, logit-function, hidden activations, and parameters therefore collapse to different degrees after grokking.

The logit-function PR of $\approx 6$ is robust across input subsets. Computing the same PR on the train set, the test set restricted to correct predictions, and the test set restricted to incorrect predictions yields 6.4, 6.3, and 6.7 respectively, all within the same low-dimensional regime. The function family's effective dimensionality is a property of the population, not an artifact of which inputs we evaluate on.

### Dispersion is not an alignment artifact

A skeptical reading of the parameter PR is that the dispersion is residual from imperfect Hungarian alignment. To rule this out, we re-computed parameter PR on subsets of models that share a single random seed (varying weight decay and learning rate within that seed). Within-seed parameter PR ranges from 15.5 to 27.3 across the five original-batch seeds, out of grokked-population sizes of 17 to 31. Each seed's PR reaches $91-97\%$ of its maximum available within-seed rank ($N-1$ for a within-seed population of $N$ grokked models). The within-seed population is itself high-dimensional; the cross-seed population's higher PR reflects genuine population spread on top of within-seed variation, not residual misalignment.

Hidden activation PR shows analogous behavior: alignment improves the cross-population cosine similarity from 0.31 to 0.52 but does not collapse PR, confirming the dispersion is not a coordinate-frame artifact.

### What the ladder tells us

The grokking phase transition collapses test accuracy. It does not collapse the model's logit-function, its hidden activations, or its parameters to a single point or a single direction. Across the 117 original-batch grokked models, the logit-function family occupies an effective subspace of $\approx 6$ directions, and the network's hidden and parameter representations spread across a much larger number of directions in their respective ambient spaces.

This raises the question that the rest of the paper investigates. The shared Fourier scaffold of §5 is consistent with hidden activation PR $\approx 74$ and parameter PR $\approx 95-110$: the algorithm is shared but its *implementation* varies along many directions. What about the logit-function PR of $\approx 6$? Six is much smaller than the hidden-activation and parameter dimensions, but it is not 1. The function family has low effective dimension but is not a point. The remainder of the paper asks: *what coordinates parameterize this low-dimensional logit-function family across initialization basins?*

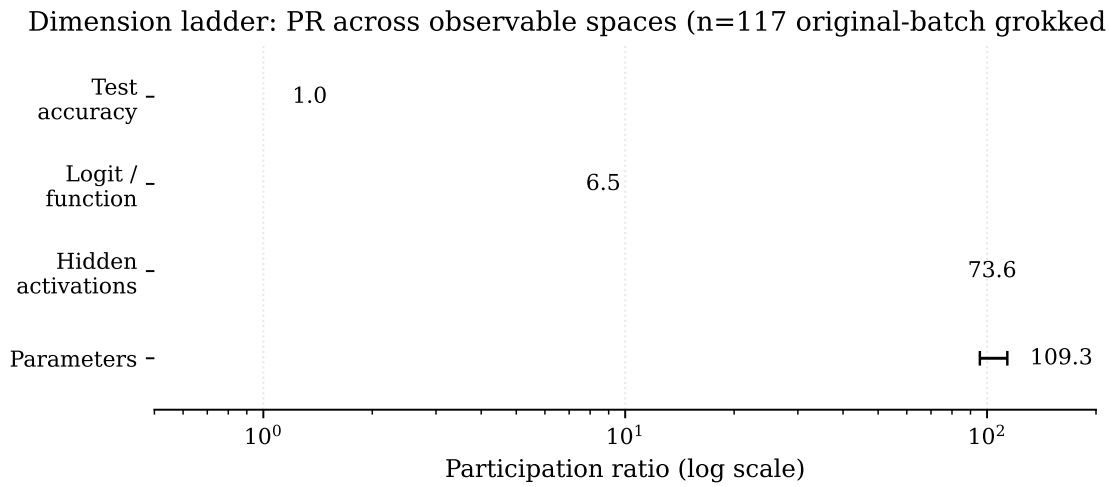


Figure 3: *The dimension ladder.* Participation ratio across four observable spaces, computed across the 117 grokked models of the original training batch (App A.5). Test accuracy is collapsed under the grokking filter; logit-function variation occupies $\approx 6$ effective directions; hidden activation variation occupies $\approx 74$; parameter variation occupies $\approx 95-110$. The four PR values span more than an order of magnitude on a log scale.

## Hidden Fourier amplitudes explain seed-local geometry but fail cross-seed

§5 established that grokked networks share a Fourier scaffold in their hidden layer. §6 established that the logit-function family these networks realize occupies an effective subspace of $\approx 6$ directions, distinct from the much higher-dimensional spread of their hidden activations and parameters. The natural hypothesis is that the hidden Fourier amplitudes — the canonical mechanistic descriptors in prior work on grokking modular addition — also serve as the cross-seed coordinates of the function family. That is, if two grokked models implement different functions within the shared

Fourier algorithm, perhaps the difference is captured by their relative Fourier amplitudes at the dominant frequencies. We test this hypothesis directly and find that it fails under held-out cross-seed evaluation.

### Selecting the hidden Fourier feature set

We construct a feature vector per model from the hidden-activation Fourier transform. For each model and each cell $(k_a, k_b)$ on the input frequency grid we compute the mean squared magnitude across hidden units:

$$F[k_a, k_b] = \frac{1}{h} \sum_{j=1}^{h} \left| \widehat{H}[k_a, k_b, j] \right|^2 .$$

We restrict to cells along the modular-addition signature lines identified in §5 — the cross lines, the diagonals, and their dominant frequencies $k \in \{4, 9, 11\}$ together with their conjugates — yielding 24 features per model. This selection follows the prior literature: these are the cells where grokked modular-addition networks concentrate Fourier mass.

### In-sample geometry recovery succeeds

We first ask whether the 24-cell feature vector reproduces the function-PC geometry of the population *without* held-out evaluation. Pooling all 217 grokked models, we fit a ridge decoder from the 24 Fourier features to the six-dimensional logit-function-PC coordinates, then compute the Mantel correlation between predicted-coordinate pairwise distances and true function-PC pairwise distances. We obtain:

$$\text{Mantel} = 0.65, \qquad R^2 = 0.66.$$

Both numbers are substantial. Reading them in isolation, one would conclude that the 24-cell feature vector is an adequate descriptor of the function family.

### In-sample $R^2$ and pairwise geometry can diverge

A diagnostic check warns that the in-sample success is more fragile than it looks. We compared the 24-cell selected Fourier feature set against a broader feature set that uses the top Fourier cells by total power without restricting to the algorithmic signature. The broader feature set achieves higher $R^2$ (0.78) but lower Mantel correlation (0.42). The top-power feature set is *better at predicting function-PC coordinates* but *worse at preserving the pairwise distance structure of the population.*

This divergence is the methodological warning we carry into §8. $R^2$ measures how well the decoder fits its own training labels; Mantel correlation measures whether the decoded feature representation preserves the actual pairwise distances in the function family. The two can come apart, especially for high-variance feature sets that overfit to leading principal components without respecting their relative structure. Throughout the paper we treat Mantel-against-shuffle as a stricter test of geometric transfer, supported by but not equivalent to $R^2$.

### Augmentations of the hidden Fourier feature set (in-sample only)

We tested whether the 24-cell amplitude vector might be missing structurally important information. Three augmentations were tried under the pooled in-sample protocol of §7.2:

- **Phase-sensitive features**: include real and imaginary components, or phase-derived summaries, to test whether sign/phase information missing from per-cell power explains the residual.
- **Coherence**: include the inner products between the complex Fourier coefficients across hidden units, capturing the spatial structure of the algorithm rather than just per-cell magnitudes.
- **Pairwise interactions**: include products $F[k_a, k_b] \cdot F[k'_a, k'_b]$ between dominant frequency cells, capturing second-order interactions.

Phase-sensitive features and coherence produced negligible Mantel improvements (within $\pm 0.02$ of the 24-cell baseline). Pairwise interactions did not substantially improve the in-sample Mantel beyond the baseline either. None of the three augmentations changed the in-sample picture. We did not evaluate these augmentations under the held-out cross-seed protocol; the cross-seed test is restricted to the two hidden-Fourier variants reported in §7.5.

### Leave-one-seed-out evaluation collapses the signal

We applied the leave-one-seed-out protocol described in §3.7 to the same 24-cell feature vector. Across 10 folds with 100-shuffle null per fold, the held-out Mantel gap is

$$0.10 \pm 0.11 \text{ at } z = 0.71.$$

Compared against the in-sample value of 0.65, this is a near-complete loss of transfer signal. The gap against shuffle is below the *meaningful* threshold (0.15) and the $z$-score is below 1. The 24-cell hidden Fourier feature vector does not transfer across initialization basins.

We additionally evaluated the broader full top-power Fourier variant from §7.3 under the same leave-one-seed-out protocol. Its held-out gap is $0.09 \pm 0.06$ at $z = 0.77$, also below the meaningful threshold. Both hidden-Fourier variants tested cross-seed — the selected 24-cell signature and the broader top-power set — fail the cross-seed transfer test, despite the broader variant's higher in-sample $R^2$. The hidden-layer Fourier description is sufficient to identify the algorithm but insufficient to identify the realized function within that algorithm across initialization basins.

### What this section establishes

The methodological lesson is that a feature set can preserve geometry within a mixed population while failing to identify transferable coordinates across initialization basins. This is not a result about Fourier features specifically; it is a result about within-population versus cross-population structure. A decoder trained and evaluated on the same population can exploit any consistent variation, including variation that is generated by within-population factors (training trajectory, schedule, weight-decay strength) and that does not correspond to the cross-population *mechanism* the experimenter is trying to identify. Held-out cross-seed evaluation is the test that separates these two.

The substantive lesson is that the cross-seed coordinates of the dialect are not located in the per-cell Fourier amplitudes of the hidden layer. The grokked Fourier scaffold of §5 is shared *as algorithm* without specifying the model's position in the dialect. The next section finds the cross-seed coordinates in the readout layer.

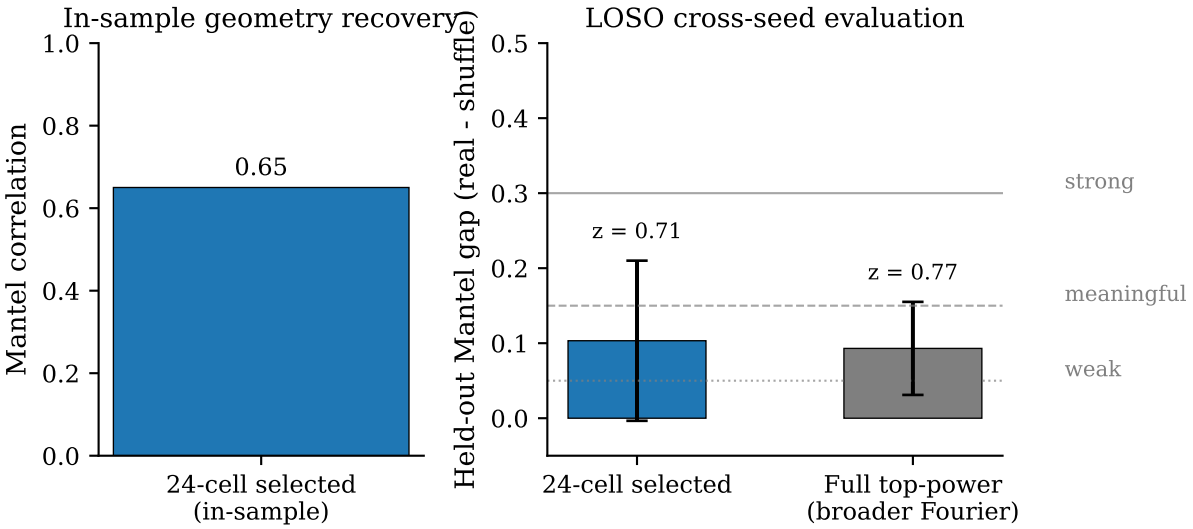


Figure 4: *In-sample success and cross-seed failure of hidden Fourier features.* (Left) In-sample 24-cell selected Fourier feature decoder: Mantel correlation 0.65 pooled across 217 grokked models. (Right) Held-out Mantel gap (real − shuffle, ±1 standard deviation across folds) under leave-one-seed-out evaluation for two hidden-Fourier feature sets: the selected 24-cell signature (gap $0.10 \pm 0.11$, $z = 0.71$) and the broader full top-power variant from §7.3 (gap $0.09 \pm 0.06$, $z = 0.77$). Horizontal lines mark the *weak* (0.05), *meaningful* (0.15), and *strong* (0.30) thresholds.

## Cross-seed function geometry is recoverable from the readout layer

The previous section showed that hidden-layer Fourier amplitudes — the canonical descriptors of the modular-addition algorithm in prior work — capture geometry within a population of grokked models but fail to identify transferable coordinates across initialization basins. We now move the search to the readout layer.

We apply the same leave-one-seed-out protocol as §7: 10 folds, one held-out seed per fold, 100 shuffled-target permutations per fold to construct a null distribution of Mantel correlations. Within each fold we fit a ridge decoder from features to the held-out seed's logit-function-PC coordinates and measure the Mantel correlation between predicted and true pairwise distances. The reported held-out gap is the difference between the real-target Mantel and the mean of the shuffled-target Mantel; the $z$-score is that gap divided by the shuffled-target standard deviation.

**Full readout succeeds where Fourier failed.** Using the flattened readout matrix $R \in \mathbb{R}^{p \times h}$ (6,016 features per model) as the feature vector, the held-out gap reaches $0.41 \pm 0.09$ at $z = 2.03$ — substantially above the $0.10 \pm 0.11$ at $z = 0.71$ obtained for selected hidden Fourier features in §7. The decoder preserves held-out geometry across initialization basins. Cross-seed function geometry,

which the hidden layer's Fourier amplitudes did not transfer, is recoverable from the readout layer.

**The signal is not feature-count advantage.** The flattened readout has 6,016 features; the hidden Fourier feature set in §7 had 24. To rule out a trivial reading in which more features simply produce better fits, we tested readout features at matched and reduced dimensionality. PCA on the flattened readout, fit per training fold, retains decoding strength at 24 components ($0.41 \pm 0.08$, $z = 2.29$) and improves at 64 components ($0.53 \pm 0.13$, $z = 2.98$). Twenty-four readout PCA components match the full 6,016-dimensional readout layer; sixty-four give the strongest mean recovery among the variants tested. The tested signal is compact.

Table 1: Held-out Mantel gap and $z$-score by feature variant under leave-one-seed-out evaluation. The hidden-Fourier variant from §7 fails; readout-derived variants succeed at matched and reduced dimensionality, ruling out feature-count advantage as the explanation.

| Variant | Features | Held-out Mantel gap | $z$ |
|---|---|---|---|
| Hidden Fourier (selected, from §7) | 24 | $0.10 \pm 0.11$ | 0.71 |
| Full $R$ (flattened readout) | 6,016 | $0.41 \pm 0.09$ | 2.03 |
| PCA($R$, 24) | 24 | $0.41 \pm 0.08$ | 2.29 |
| PCA($R$, 64) | 64 | $\mathbf{0.53 \pm 0.13}$ | **2.98** |

The PCA-24 variant is particularly diagnostic. With the *same* feature count as the failing hidden-Fourier variant in §7, the readout PCA produces a roughly four-fold larger gap and a $z$-score above two standard deviations; PCA-64 approaches a three-standard-deviation effect. The contrast isolates *where* the cross-seed signal lives, not *how much information* the feature vector contains.

**Coordinate prediction is weak; geometry recovery is what we claim.** Across folds, per-PC $R^2$ is heterogeneous and often negative, indicating that the decoder more reliably preserves pairwise geometry than exact held-out PC-coordinate values. Throughout this paper we report Mantel-gap-against-shuffle as the held-out test of cross-seed transfer. Coordinate prediction is the stricter goal we do not claim.

**The most consistently coordinate-predictable PCs are middle PCs.** Examining the per-PC $R^2$ heatmap of the readout-feature decoder fold by fold (Figure 5) reveals a non-uniform distribution of recovery across the six function PCs. PCs 2 through 4 show consistently positive $R^2$ across folds; PCs 1, 5, and 6 show consistently negative or near-zero $R^2$. The leading function-PC, which carries the largest fraction of logit variance in the population, is the dimension along which the decoder fails most. Middle function-PCs, which carry less variance, are where the decoder's predictions track the held-out data.

This finding has two implications. First, the flat-statistic reading "the function family has effective dimension $\approx 6$, all six dimensions transfer equally" is wrong; the most consistently coordinate-predictable directions in this test are PC2 through PC4, not the full leading-six logit-function subspace. Second, the leading function-PC carries high variance, but that variance is not predictable from any of our tested feature sets. Whether PC1 represents a label-permutation gauge, a multi-frequency aggregate that no single readout summary captures, or something else, is a question we return to in §9.4 in the intervention setting.

**Summary of §8.** The cross-seed function geometry, which hidden-layer Fourier amplitudes failed to identify, is recoverable from the readout layer. The recoverable signal is compact (PCA-24 already matches the full readout, PCA-64 produces the strongest tested recovery), and the most consistently coordinate-predictable function-PCs are middle PCs (PC2–PC4) rather than the dominant PC1. These findings establish the *location* and *coarse dimensionality* of the dialect subspace but leave its interpretation opaque: PCA components of a readout matrix have no inherent algorithmic meaning. §9 names the basis.

---

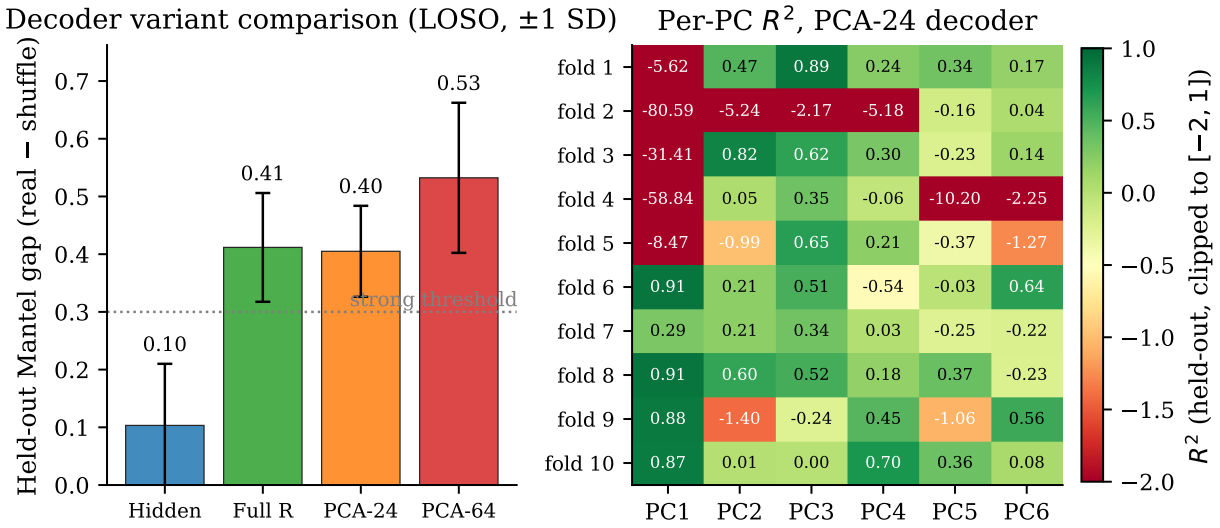


Figure 5: *Cross-seed decoding succeeds at the readout layer.* (Left) Held-out Mantel gap (real-target minus shuffled-target mean) by feature variant: hidden Fourier features from §7, full flattened readout, and two PCA-readout variants. Error bars span ±1 standard deviation across folds. The horizontal dotted line at 0.30 marks the *strong* threshold used throughout this paper. (Right) Per-PC $R^2$ heatmap for the readout-PCA-24 decoder, fold by fold, on the six logit-function PCs of the held-out seed. Positive values (green) indicate the decoder explains variance on that PC; negative values (red) indicate failure. The middle PCs (PC2–PC4) are most consistently coordinate-predictable; the leading PC1 and the smaller PC5–PC6 are not. Color scale clipped to $[-2, 1]$; cell annotations show unclipped $R^2$.

## The readout subspace is a gauge-invariant irrep-energy budget

The previous section established that the dialect subspace is recoverable from a compact subspace of the readout layer, but the recovered subspace is described in PCA coordinates whose interpretation is opaque. In this section we name the subspace. We show that much of the transferable dialect geometry is captured by the per-irrep energy distribution of the readout matrix — a gauge-invariant 47-dimensional vector that collapses by exact Hermitian symmetry to 24 independent dimensions, brackets to an effective dimension between 12 and 24 by PCA, and can be manipulated to move models along predicted directions in the transferable function subspace. The irrep-energy budget is the coordinate system we use to parameterize the dialect.

### An irrep-energy decoder of the readout layer

Modular addition over $\mathbb{Z}/p\mathbb{Z}$ has a natural Fourier decomposition: the irreducible representations of the cyclic group are indexed by frequencies $k \in \{0, 1, \ldots, p-1\}$, and any function $f : \mathbb{Z}/p\mathbb{Z} \to \mathbb{C}$ admits a discrete Fourier expansion in this basis. Prior work (Nanda et al. 2023) established that grokked networks discover this Fourier structure in their hidden representations. We extend the analysis to the readout layer.

Let $R \in \mathbb{R}^{p \times h}$ denote the aligned readout matrix (alignment described in §3.5), where $p = 47$ is the modulus and $h = 128$ is the hidden width. Each column of $R$ is a function from output index $y \in \{0, \ldots, p-1\}$ to a scalar weight. We take the discrete Fourier transform of $R$ along the output axis:

$$\widehat{R}[k, j] = \sum_{y=0}^{p-1} R[y, j]\, e^{-2\pi i k\, y/p}, \qquad k \in \{0, \ldots, p-1\}, \quad j \in \{1, \ldots, h\}.$$

The per-irrep readout energy of a model is then the squared magnitude summed across hidden units:

$$M[k] = \sum_{j=1}^{h} \left|\widehat{R}[k, j]\right|^2.$$

This produces a single 47-dimensional vector $M \in \mathbb{R}^{47}$ per model. Each entry is the total readout power that model allocates to the corresponding irrep of $\mathbb{Z}/47\mathbb{Z}$. Because the sum over hidden units integrates out per-unit assignments, $M$ is invariant under any unit-permutation gauge.

We evaluated $M$ as a feature for held-out cross-seed function-geometry decoding using the leave-one-seed-out protocol described in §3.7 (10 folds, 100-shuffle null per fold). The 47-dimensional irrep-energy vector matches the full readout and readout-PCA variants within error bars while providing an interpretable basis:

@ >

p(- 6) * 0.2500 >

p(- 6) * 0.2500 >

p(- 6) * 0.2500 >

p(- 6) * 0.2500@

Variant

&

Features

&

Mantel gap (real − shuffle)

&

$z$

Full $R$ (flattened readout, $p \times h = 6{,}016$) & 6,016 & $0.41 \pm 0.09$ & 2.03
PCA($R$, 24) & 24 & $0.41 \pm 0.08$ & 2.29
PCA($R$, 64) & 64 & $0.53 \pm 0.13$ & 2.98
$M$ **(irrep energies)** & **47** & $\mathbf{0.47 \pm 0.16}$ & **2.94**

The structural choice — Fourier basis along the output axis, summed across hidden units — is principled rather than learned, and it produces a representation that is interpretable in terms of the shared Fourier scaffold.

### Hermitian symmetry and the effective dimension

The readout matrix $R$ is real-valued, so its Fourier transform along the output axis satisfies the Hermitian symmetry $\widehat{R}[k, j] = \overline{\widehat{R}[p-k, j]}$, and consequently $|\widehat{R}[k, j]|^2 = |\widehat{R}[p-k, j]|^2$ pointwise. We verified this empirically: the ratio of total power between conjugate-collapsed and uncollapsed representations is 1.0000 to four significant figures across all 217 models. The 47-dimensional irrep-energy vector therefore contains 24 independent real-valued dimensions: $M[0]$ (DC), and 23 conjugate-pair sums $M[k] + M[p-k]$ for $k = 1, \ldots, 23$.

Decoding survives the collapse. The 24-dimensional Hermitian-collapsed vector $M^{(24)}$ achieves a held-out gap of $0.43 \pm 0.13$ at $z = 2.50$ — within error bars of the full 47-dimensional version. Thus the Hermitian collapse halves the feature dimension without materially reducing decoding performance.

We further bracketed the effective dimension by PCA on $M$ within each training fold. PCA-12 retains a meaningful gap ($0.37 \pm 0.14$, $z = 1.99$); PCA-6 falls below the meaningful threshold ($0.20 \pm 0.08$, $z = 1.21$). The transferable signal lies in an effective subspace whose dimension is bracketed by $12 \leq d \leq 24$. We do not claim the dimension exactly; we claim the bracket.

This dimensional bracket converges with the independent finding from §8 that PCA on the raw readout matrix reaches its strongest decoding at 24–64 components. The two analyses point at the same compact subspace from different bases.

### Per-unit amplitudes and phases as gauge

The aggregation across hidden units is essential. To see why, we examined the per-cell statistics of $\widehat{R}[k, j]$ across the 217 grokked models.

Per-cell amplitudes vary substantially across initializations: at the dominant frequencies $k \in \{4, 11\}$ the coefficient of variation $\mathrm{std}(|\widehat{R}|)/\mathrm{mean}(|\widehat{R}|)$ averaged across hidden units is in the range 1.5–1.9. Per-cell phases are essentially uniform: the circular concentration $|\langle e^{i \angle \widehat{R}[k,j]} \rangle|$ averaged across hidden units sits at approximately 0.17 for every frequency, close to the value expected for phases drawn uniformly at random.

Per-cell amplitudes and phases behave like gauge-dependent coordinates of the function $R$ implements. Different initialization basins can arrive at different unit-level decompositions, while the

aggregate $M$ remains the transferable coordinate. The aggregate — which integrates these per-unit choices out — is what carries the cross-seed signal.

This is consistent with our observation in §7 that hidden-layer Fourier amplitudes also fail cross-seed transfer. Per-unit amplitudes are part of the gauge-dependent description at every layer the network exposes; only the appropriately aggregated quantity transfers.

If $M$ is merely a better summary statistic, interventions on it need not move the model's logit-function geometry. The next test asks whether changing a model's irrep budget directly produces the function-space shifts predicted by the population-level decoder.

### Intervention on dominant and control irreps

The decoding evidence shows that $M$ is correlated with logit-function variation across initialization basins. Correlation does not establish that $M$ is a causally accessible coordinate of the function family. To probe this we performed surgical interventions on the irrep budget and measured whether logit-function geometry moves in the predicted direction.

**Procedure.** For each test model we modified the aligned readout matrix in the irrep basis. We selected a target frequency $k$, a scaling factor $s$, and constructed a modified Fourier representation by setting

$$\widehat{R}'[k, j] = s \cdot \widehat{R}[k, j], \qquad \widehat{R}'[p-k, j] = s \cdot \widehat{R}[p-k, j],$$

preserving the Hermitian conjugate to keep the inverse-transformed weights real-valued, and leaving all other frequencies unchanged. We then reconstructed $R'$ via inverse DFT, replaced the model's readout, and ran a forward pass on the test inputs to obtain new logits. The logits were projected onto the function-PC basis computed from the full 217-model population, yielding an observed shift $\Delta z_{\text{obs}} \in \mathbb{R}^6$. A linear ridge decoder $M \to z$ fitted on the same full-population baseline provided a predicted shift $\Delta z_{\text{pred}}$ from the change $\Delta M$ induced by the intervention. We compared predicted and observed shifts.

The intervention test uses a population-level decoder rather than a leave-one-out decoder. Each intervened model is one of 217 models in the decoder's training set. The intervention does not test whether the decoder generalizes to *unseen* models — that question is answered by the cross-seed evaluation in §8. The intervention test asks a different question: *given the population's learned $M$-to-function mapping, do counterfactual changes to $M$ produce the function shifts the mapping predicts?* This is the appropriate framing for a partial causal claim about the irrep budget: we are not asking whether $M$ predicts function in new data, we are asking whether counterfactual changes to $M$ move function in the predicted directions in this system.

We tested 20 grokked models, three frequencies, and five scaling factors $s \in \{0, 0.5, 0.75, 1.5, 2.0\}$, for a total of 300 interventions. Two of the tested frequencies are dominant in the population mean spectrum ($k = 4$ and $k = 11$, ranked first and second by mean readout power among the conjugate pairs). The third is a low-power control ($k = 23$, ranked near the bottom of the spectrum).

**Sanity: grokking is preserved.** Across all 300 interventions the resulting test accuracy remained at or near the grokked baseline. At dominant frequencies with $s = 0$ (full ablation of the irrep), 19 of 20 models retained acc $\geq 0.95$ at $k = 4$ and 18 of 20 at $k = 11$; at all moderate scales ($s \in \{0.5, 0.75, 1.5, 2.0\}$), all 20 models retained acc $\geq 0.95$. At the control frequency $k = 23$, all

20 models retained accuracy across all scales. Surgical readout intervention moves models within the grokked function family rather than off it, validating the manipulation as a probe of function geometry rather than network failure.

**Result: intervention moves logit-function geometry in the predicted direction at dominant frequencies; control does not.**

We computed direction agreement between predicted and observed shifts on the transferable subspace identified in §8 (PCs 2 through 6, where cross-seed decoding succeeded above the leading PC1) using cosine similarity. We use the broader PC2–PC6 subspace for the intervention test because PC1 was independently identified in §8 as non-transferable, leaving PC2 through PC6 as the non-leading logit-function subspace; the strongest coordinate-prediction evidence in §8 lies more narrowly within PC2–PC4, but the intervention test is the appropriate place to evaluate the broader transferable region.

Table 2: Direction agreement between predicted and observed function-space shifts on PC2–PC6, by intervention frequency. Dominant frequencies ($k = 4, 11$) produce predicted-direction agreement; the low-power control ($k = 23$) does not.

| Frequency | Role | $n$ | mean cos | median cos | fraction $> 0.5$ |
|---|---|---|---|---|---|
| $k = 4$ | dominant | 100 | $+0.62$ | $+0.98$ | 0.80 |
| $k = 11$ | dominant | 100 | $+0.32$ | $+0.99$ | 0.65 |
| $k = 23$ | control | 100 | $-0.23$ | $-0.16$ | 0.00 |

Per-PC Pearson correlation between predicted and observed shifts on PC2–PC6 reaches $r = +0.918$ at $k = 4$ and $r = +0.893$ at $k = 11$. Across all 300 interventions, the magnitude correlation is $r = +0.77$, with observed shifts systematically smaller than predicted (regression slope below unity).

The control frequency does not produce predictable function shifts. This argues against a generic-perturbation explanation in which arbitrary modifications of $R$ produce correlated movement: in our test, only modifications at frequencies where the population concentrates readout mass move function as the budget-decoder predicts.

**Caveats.**

*PC1 is not reliably governed by the tested intervention.* Across the six function-PC dimensions, PC1 shows frequency-dependent signed responses that the per-frequency decoder does not consistently capture (per-trial cosines on PC1-only are sign-flippy: $-0.70$ mean at $k = 4$, $+0.80$ at $k = 11$). A 1-D cosine collapses to $\pm 1$ per intervention, so the per-trial values are themselves brittle. We do not interpret PC1 as proven gauge; we report it as not under reliable causal control by the tested per-frequency manipulation, and we restrict our intervention claims to the transferable subspace PC2–PC6 identified independently in §8. The structure of PC1 — whether it represents a label-permutation gauge, a multi-frequency balance the per-frequency decoder cannot capture, or something else — is left as an open question.

$k = 11$ *shows bimodality.* The $k = 11$ direction-agreement distribution has a median near $+0.99$ but a substantial anti-aligned tail, producing a mean of only $+0.32$. Most interventions at $k = 11$ produce near-perfect predicted-direction agreement; a minority produce strong anti-alignment. This

is consistent with subpopulation structure within the 217 grokked models — distinct init-basin attractors with different signed couplings between $k = 11$ budget and PC2–PC6 — but identification of which models fall into which subgroup is beyond the scope of this paper.

*Magnitudes are damped.* Observed shifts are correlated with predicted shifts but systematically smaller. This is consistent with the irrep-energy budget being a coupled property of the full network rather than a property localized in the readout: the rest of the network — the embedding, the first layer, the nonlinearity — partially attenuates the perturbation, since they continue to compute the same hidden representation while only the readout has been modified. Full causal identification would require coupled-layer intervention, which we leave to future work. We claim partial causal traction on the transferable subspace, not complete causal identification of the budget.

**Summary.** Surgical intervention on the irrep budget at frequencies where the population concentrates its readout mass moves function in directions predicted by the budget-to-function decoder, on the transferable subspace, while preserving grokking. The same intervention at a control frequency produces no systematic effect. This pattern is what an intervention experiment should produce if the irrep budget is a real coordinate of the dialect rather than a correlational artifact of the decoder's training population. It is not what a complete causal theory would produce — magnitudes are damped, PC1 is not under reliable control, and one of the dominant frequencies shows bimodal coupling — but it is enough to upgrade the irrep budget from a decoded coordinate to a causally accessible one.

*The irrep budget is not just decoded from the dialect; it is a partially controllable coordinate of it.*

## The dialect direction is unusually function-preserving under parameter perturbation

§9.4 showed that intervention on the irrep budget at dominant frequencies moves logit-function geometry in the directions the population's decoder predicts, while a control-frequency intervention produces no systematic effect. That experiment establishes one direction (the dialect direction, defined by the readout-Fourier intervention) and one comparison (a control frequency). We now ask the more general question: does the dialect direction differ *categorically* from arbitrary directions in parameter space, or only from one specific control? If the dialect subspace is causally meaningful, perturbations along it should preserve function while perturbations along generic parameter directions should not.

**Procedure.** We construct four direction families in the flat parameter space $\mathbb{R}^{15{,}887}$ of the network, plus a fifth as a sanity check.

1. *Dialect direction.* For each model we compute the unit vector in parameter space corresponding to the readout-Fourier intervention from §9.4, normalized to unit L2 length. Concretely: scale the readout's Fourier coefficients at all dominant frequencies by 1.5 (preserving Hermitian symmetry), inverse-DFT to obtain a perturbed readout, and use the difference vector — embedded into the full parameter vector with zeros for embedding and fc1 entries — as the direction.

2. *Scaffold-sensitive direction.* The gradient of dominant-frequency Fourier power on hidden activations with respect to all parameters: $\nabla_\theta \sum_{k \in K} \log \left|\widehat{H}[\cdot, k, \cdot]\right|^2$ over the dominant-frequency cells $K$ of the hidden layer's 2D DFT. Since the scaffold lives in hidden activations, this gradient is identically zero on fc2 parameters and concentrates on embedding and fc1 — a

geometric reflection of the §9 observation that the scaffold is a hidden-layer property and the dialect is a readout-layer property. We normalize to unit L2 length.

3. *Random direction (full parameter).* A Gaussian random vector in flat parameter space, normalized to unit L2 length. Five fresh samples per perturbation norm are averaged within model.

4. *fc2-random direction (locality-matched control).* A Gaussian random vector restricted to the 6,063 entries of the flat parameter vector that belong to fc2.weight and fc2.bias, with zeros elsewhere, normalized to unit L2 length. This controls for the alternative hypothesis that the dialect direction's function-preservation is explained merely by being localized to the readout layer — which contains 38% of the parameter budget — rather than by the readout-Fourier structure. Five fresh samples per perturbation norm are averaged within model.

5. *Parameter-symmetry sanity check.* Apply hidden-unit permutation jointly to fc1 rows and fc2 columns, leaving the embedding unchanged. This is an exact functional invariance: the resulting model implements the same function as the original.

For each model and each direction family we sweep the perturbation L2 norm $\|\Delta\theta\|$ across 18 log-spaced values from $10^{-2}$ to $10^{2}$, replace the model's parameters by $\theta + \|\Delta\theta\| \cdot \hat{d}$, and measure the resulting test accuracy on the held-out test split.

**Result: the dialect direction preserves function across four orders of magnitude.** Across all 217 grokked models, median test accuracy along the dialect direction remains $\geq 0.97$ across the entire $10^{-2}$ to $10^{2}$ sweep. The three control directions cluster together at much lower preservation: median accuracy crosses below the grokking threshold 0.95 at $\|\Delta\theta\| \approx 8.0$ for scaffold-sensitive, $\|\Delta\theta\| \approx 11.8$ for fc2-random, and $\|\Delta\theta\| \approx 12.4$ for full-parameter random. By $\|\Delta\theta\| = 100$ all three collapse far below the grokking threshold, with full-parameter random (0.024) and scaffold-sensitive (0.036) near chance ($1/p \approx 0.021$) and fc2-random at 0.080 — distinctly above chance but well below grokking. The dialect direction holds at 0.969.

@ >

p(- 4) * 0.3333 >

p(- 4) * 0.3333 >

p(- 4) * 0.3333@

Direction family

&

$\|\Delta\theta\|$ at first crossing of 0.95

&

Median acc at $\|\Delta\theta\| = 100$

Dialect & not crossed & 0.969
fc2-random (locality control) & 11.8 & 0.080
Full-parameter random & 12.4 & 0.024

Scaffold-sensitive & 8.0 & 0.036
Parameter-symmetry sanity check & not applicable & mean $\|\Delta\text{acc}\| < 10^{-6}$

For reference, chance accuracy is $1/p \approx 0.021$. Full-parameter random and scaffold-sensitive directions reach chance by $\|\Delta\theta\| = 100$; fc2-random reaches 0.080, distinctly above chance but well below the grokking threshold.

The parameter-symmetry sanity check confirms the experiment is correctly constructed: applying an exact functional invariance preserves test accuracy to numerical precision. The interpretation is that the experiment can detect zero functional change when zero is expected, supporting the substantive separation observed between the dialect direction and the other three families.

**The fc2-locality control is the key disambiguation.** A skeptical reading of §9.4 might attribute the dialect direction's function-preservation to its being localized to fc2 — the layer with 38% of the parameter budget, where small per-coordinate perturbations might be expected to do less damage than perturbations spread across the whole network. The fc2-random comparison directly tests this: a unit vector restricted to fc2 entries breaks the function at $\|\Delta\theta\| \approx 11.8$, statistically indistinguishable from full-parameter random at $\|\Delta\theta\| \approx 12.4$. fc2-locality alone does not explain dialect's function-preservation. The dialect direction's robustness is a property of the *readout-Fourier structure* of the perturbation, not its parameter-space support.

**Interpretation.** Across four orders of magnitude in perturbation norm, the dialect direction is *sharply separated* from generic parameter perturbation. The dialect direction tested here is not a statistical artifact of the cross-seed decoder — it is empirically a function-preserving direction in the local geometry of parameter space. The three control directions cluster tightly at the grokking-threshold crossing (norm $\approx$ 8–12), while dialect remains preserved an order of magnitude further. Critically, fc2-locality is ruled out as the explanatory variable: structured readout-Fourier modification, not readout-layer locality, is what makes the dialect direction function-preserving.

This result tightens the boundary statement of §9.4 from "partial causal traction on the transferable subspace" to "measurable causal asymmetry under parameter perturbation between the dialect direction and other directions in parameter space." We do not claim complete causal identification of the dialect's internal coordinate structure — PC1 remains the open question it was in §9.4 — and we do not claim that arbitrary points in the dialect subspace are reachable by intervention on the irrep budget alone. §11 returns to these distinctions.

### Schedule effects in irrep-budget space (coda)

A subset of our 217 models were trained with cyclic weight-decay schedules (44 models) rather than constant weight decay (173 models). The cyclic-schedule subpopulation occupies a measurably distinct region of irrep-budget space: the Mahalanobis distance between cyclic and constant centroids in $M^{(47)}$ is 1.91, and the same distance in $M^{(24)}$ is also 1.91.[2] Training schedule therefore appears as a shift in irrep-budget space, suggesting schedule structure affects where models land in dialect space. This is supporting evidence rather than the central mechanism; the dialect coordinate identified in §9.1–§9.5 is the load-bearing claim.

[2]We omit raw-readout-space Mahalanobis from the main result because the rank-deficient covariance matrix produces pseudo-inverse instability and a numerically misleading large value. PCA-projected variants $M_{\text{PCA-6}}$ (0.63) and $M_{\text{PCA-12}}$ (0.96) are reported in the appendix.

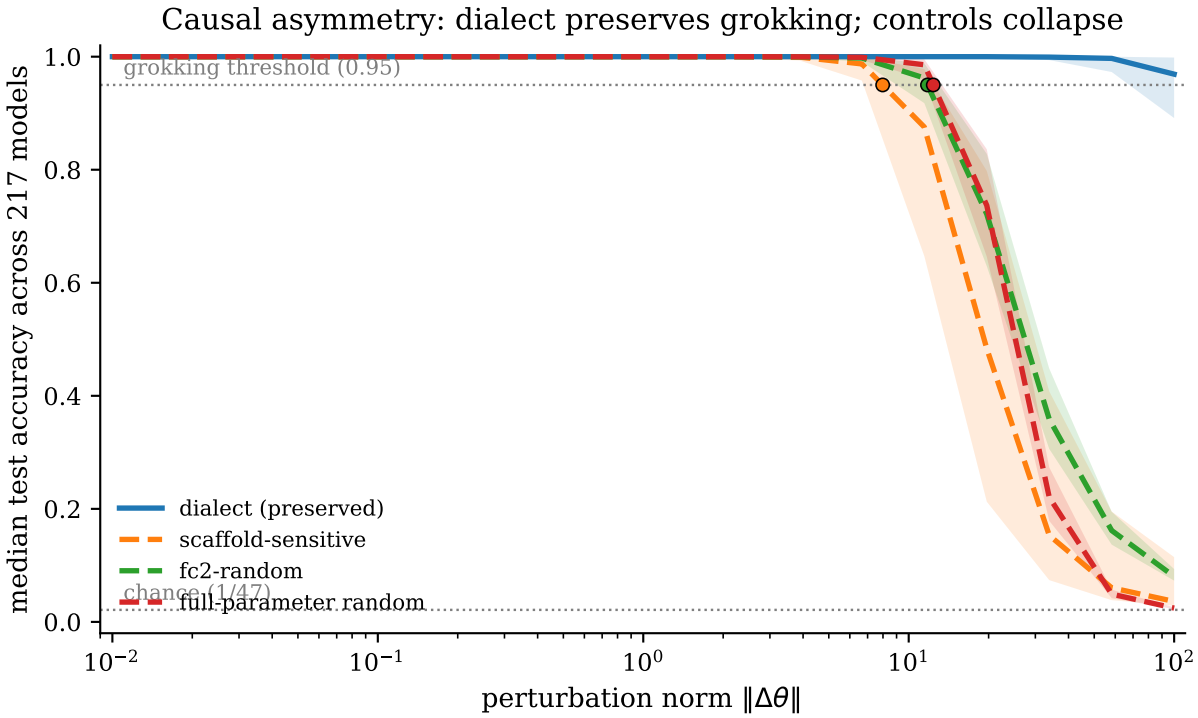


Figure 6: *Causal asymmetry of the dialect subspace.* Median test accuracy across the 217 grokked models for four direction families perturbed at a sweep of L2 norms; bands show 25th–75th percentile across models. The dialect direction (solid, dark) preserves grokking across four orders of magnitude in perturbation norm; scaffold-sensitive ($\|\Delta\theta\| \approx 8.0$), fc2-random ($\approx 11.8$), and full-parameter random ($\approx 12.4$) cross below the grokking threshold in the 8–12 range and collapse far below threshold by $\|\Delta\theta\| = 100$ (full-random and scaffold-sensitive at chance, fc2-random at 0.080). fc2-random tracks full-parameter random closely, ruling out fc2-locality as the explanation for dialect's function preservation. Parameter-symmetry sanity check: hidden-unit permutation preserves test accuracy to numerical precision (mean $|\Delta\text{acc}| < 10^{-6}$, $n = 217$).

# Discussion

## What we showed

Across 217 grokked modular-addition networks trained from 10 random seeds and a sweep of weight-decay strengths, learning rates, and decay schedules, four observations cohere into a single picture of post-grokking structure.

First, the analyzed grokked models concentrate their hidden-activation Fourier mass on the same algorithmic signature: cross lines and diagonals on the input-frequency grid, with dominant power at the same conjugate-pair frequencies. Grokking, given that it occurs, produces a shared Fourier *scaffold* (§5).

Second, despite this shared scaffold, the realized functions vary. In the original-batch PR analysis, the logit-function family — the *dialect* — occupies an effective subspace of $\sim 6$ directions in a 62,322-dimensional ambient space, much smaller than the ambient space but not a single point. Hidden activations and parameters spread across $\sim 74$ and $\sim 95-110$ effective directions respectively in the original-batch subpopulation (§6). The dispersion is real and structured, not within-seed noise.

Third, the cross-seed coordinates of the dialect are not in the hidden Fourier amplitudes that prior work uses to describe the modular-addition algorithm. Hidden-Fourier features explain population geometry within a pooled sample (Mantel 0.65) but fail leave-one-seed-out evaluation (gap 0.10, $z = 0.71$, §7). The dialect subspace is recoverable from, and partially controllable through, the readout layer: a 24-dimensional gauge-invariant aggregate — total Fourier power per irrep of $\mathbb{Z}/47\mathbb{Z}$ — recovers held-out function geometry at least as well as the full 6,016-dimensional flattened readout while being structurally interpretable (§§8–9). Surgical intervention on this aggregate at dominant frequencies moves logit-function geometry in the directions the population-level decoder predicts, on the non-leading function-PC subspace, while preserving grokking. Intervention at a low-power control frequency produces no systematic effect (§9.4).

Fourth, the dialect direction is causally asymmetric under parameter perturbation: along the dialect direction, perturbations of magnitude up to $\|\Delta\theta\| = 100$ preserve test accuracy at $\geq 0.97$, while three control families — random parameter directions, fc2-locality-matched random directions, and scaffold-sensitive directions — all break the function at perturbation norms in the 8–12 range and collapse far below threshold by $\|\Delta\theta\| = 100$ (§9.5). The fc2-locality control rules out the alternative hypothesis that dialect's robustness reflects readout-layer locality rather than readout-Fourier structure: fc2-random tracks full-parameter random closely. The dialect direction is sharply separated from generic parameter perturbation; the function family contains a function-preserving direction whose extent is measurable.

Together, these results argue that post-grokking is not a single solution but a structured family — a shared scaffold and a structured dialect within it. Models share an algorithmic scaffold — the Fourier alphabet — and differ along an interpretable readout coordinate — the irrep-energy budget — that indexes the dialect, with intervention-supported partial causal traction on its non-leading directions and measurable function-preservation along the dialect direction across four orders of magnitude in perturbation norm.

## Why it matters

The methodological recommendation is simple and concrete: **held-out cross-seed evaluation should be a default for representational-mechanism claims.** Within-population decoding

can produce strong-looking results from feature sets that fail to identify any cross-seed coordinate — our hidden-Fourier result shows this directly, with a pooled-sample Mantel of 0.65 collapsing to a held-out gap of 0.10 on the same feature set. A representational claim that survives the held-out test is a stronger claim than a representational claim that does not. We recommend that mechanistic interpretability work in this regime — particularly work that purports to identify the coordinates of a function family — report leave-one-seed-out gaps with shuffled-target nulls as a standard supplement to in-sample correlations.

The substantive recommendation is that single-checkpoint mechanistic interpretability is not sufficient for identifying population-level underdetermination. A single grokked checkpoint exhibits the Fourier scaffold of §5 and a specific irrep-energy budget; describing only the scaffold misses the budget, and describing only the budget misses that the budget varies meaningfully across initialization basins. Population-level analysis is necessary to identify what is universal versus what is gauge versus what indexes the function-level variation.

The intervention with a control frequency is a falsification structure that we recommend more broadly. Mechanism claims about feature sets often face the objection that any feature set with enough degrees of freedom will appear correlated with the target. The minimal counter is to identify, within the same feature family, a member that the mechanism predicts should not couple — and to show that intervening on it produces no effect. We did this with $k = 23$. The null result at the control intervention is what distinguishes our result from "modifying the readout produces correlated movement in something."

### Open questions

*Training-time governance.* What determines a specific irrep budget at training time? Schedule shifts the budget (Mahalanobis of 1.91 between cyclic and constant centroids in $M$-space, §9.6); initialization shifts it more. The mechanism by which initialization commits the network to one budget rather than another is the deepest open question and the one most likely to require theoretical rather than empirical work. It is also the question whose answer would most directly extend the present paper into a closed account of grokking dynamics.

*Coupled-layer intervention.* The magnitude damping in §9.4 — observed shifts smaller than predicted shifts despite strong correlation — and the function-preservation along the dialect direction in §9.5 are consistent with the dialect subspace being maintained jointly by the embedding, fc1, and fc2 layers rather than localized in the readout. The natural follow-up is intervention that preserves the irrep budget across all three layers simultaneously, which we expect would close the §9.4 magnitude gap and let us test how much of the dialect subspace — beyond the direction tested in §9.5 — is reachable by coupled intervention. This is concrete future work and the most direct path to upgrading our measurable causal asymmetry to complete causal identification of the dialect's internal coordinate structure.

*PC1 structure.* The leading function-PC carries high variance in the population but is not predictable from any of our tested feature sets, and the per-frequency intervention does not reliably govern it. PC1 may represent a label-permutation gauge that no per-frequency readout summary captures, a multi-frequency aggregate orthogonal to our irrep basis, or something else. Resolving its structure likely requires (a) coupled fc1/fc2 intervention as above, (b) embedding-layer features tested in the same protocol, or (c) a theoretical model of how training selects the budget that predicts the gauge structure of the leading PC.

*Bimodality at $k = 11$.* The intervention at $k = 11$ produces median direction agreement near +0.99 on PC2–PC6, but with a substantial anti-aligned subpopulation tail. The simplest explanation is that the 217-model population contains distinct init-basin attractors with different signed couplings between $k = 11$ budget and the transferable function subspace. Identifying which models fall into which subgroup, and what trains the difference, is open.

*Universality across tasks.* The strongest generalization suggested by our framework is that the irrep-energy mechanism is mathematically natural for cyclic-group operations: any classification head over a cyclic group output space admits a Fourier decomposition along the output axis, and the per-irrep aggregate is automatically gauge-invariant under hidden-unit permutation. Whether this mathematical naturalness predicts that the mechanism appears empirically in non-grokking regimes, in non-cyclic groups, in deeper architectures, or in tasks without explicit group structure is the most important direction for follow-up. We are pursuing replication at $p = 53$ and $p = 97$.

## Limitations

We restate and consolidate the boundaries of our claims here, in a single place, for ease of reviewer reference.

**Empirical scope.** All experiments use modular addition over $\mathbb{Z}/47\mathbb{Z}$, a two-layer MLP with hidden width 128, and AdamW with weight decay. We do not test other primes, other groups (cyclic or otherwise), other architectures (deeper networks, transformer-style attention layers, different nonlinearities), or other optimizers. The mechanism we identify — the irrep-energy budget — is mathematically natural for cyclic-group classification heads, but whether it appears empirically beyond this specific testbed is not established by this paper.

**Population size and statistical precision.** Our population is 217 grokked models drawn from 232 attempted training runs (a $\geq 0.95$ test-accuracy filter). This is sufficient for held-out cross-seed evaluation with leave-one-seed-out across 10 seeds and shuffled-target nulls with $z$-scores in the 2–3 range. It is not sufficient for tight $z$-score confidence intervals across all variants, and it is not sufficient to resolve fine subpopulation structure such as the bimodal coupling we observe at $k = 11$ (§9.4).

**Geometry recovery, not coordinate prediction.** Throughout, our cross-seed decoders preserve pairwise function-PC distances (Mantel against shuffle) on held-out seeds. They do not predict the held-out seed's exact PC-coordinate values: per-PC $R^2$ is negative on most folds. The paper's claim is geometry recovery, not coordinate prediction.

**Measurable causal asymmetry, not complete causal identification.** Intervention on the irrep budget at the readout layer (§9.4) shows magnitude correlation $r = +0.77$ between predicted and observed shifts with regression slope below unity — partial, magnitude-damped controllability. Perturbation along the dialect direction (§9.5) shows function preservation across four orders of magnitude while three control directions (random, fc2-locality-matched random, scaffold-sensitive) all break the function at perturbation norms in the 8–12 range. Together these establish causal asymmetry between the dialect direction and other parameter-space directions, and rule out fc2-locality alone as the explanation. They do not establish that arbitrary points in the dialect subspace are reachable by intervention; only that the tested direction is function-preserving. Coupled-layer intervention would close the magnitude gap of §9.4 and is left to future work.

**Local mode connectivity, not global topology.** §4 establishes that same-seed grokked end-

points under different weight-decay configurations are connected by curved low-loss paths. We do not investigate cross-seed mode connectivity, nor the topology of the full grokked manifold. The mode-connectivity claim is bounded to within-seed pairs at different weight-decay configurations.

**Conditioning on grokking.** All population statistics are computed across grokked models that reached our test-accuracy threshold. We do not characterize the irrep-energy budget for non-grokked endpoints; we make no claim about the budget along the training trajectory before grokking; and our universality language is qualified throughout as *given grokking*, not as a claim about all trained networks of this architecture.

**PC1 is not under reliable causal control.** Across the six logit-function PCs, PC1 shows frequency-coupled signed responses that the per-frequency intervention does not consistently capture. We restrict our intervention claims to PC2–PC6 and treat PC1's structure as an open question (§10.3).

**Alignment is one-shot.** Our cross-model alignment uses activation-based Hungarian matching to a single reference model. Soft matching, joint multi-target alignment, or alignment that propagates uncertainty would likely tighten the population-level dispersion measurements but is not pursued here.

# Conclusion

The grokking phase transition collapses test accuracy. It does not collapse the network. Across 217 grokked models trained on modular addition, the same Fourier algorithm appears — the *scaffold* — but the realized function family across initialization basins — the *dialect* — occupies an effective subspace of low but non-trivial dimension. The cross-seed coordinates of the dialect are not in the hidden-layer Fourier amplitudes that the algorithm's mechanistic description uses; they are recoverable from, and partially controllable through, the readout layer, in a structurally interpretable basis given by the irreducible representations of the output group. A 24-dimensional gauge-invariant aggregate — the per-irrep readout energy budget — recovers held-out function geometry at least as well as the full readout within error bars while collapsing per-unit gauge structure into a transferable coordinate system. Surgical intervention on this budget at frequencies where the population concentrates its readout mass moves logit-function geometry in directions the population's own decoder predicts, on the non-leading function-PC subspace, while preserving grokking. The same intervention at a low-power control frequency produces no such effect. Parameter-space perturbations show that this dialect direction is causally asymmetric: it preserves function across four orders of magnitude in perturbation norm while locality-matched and scaffold-sensitive controls do not. The hidden layer supplies the Fourier alphabet; the readout indexes the dialect; and the dialect direction is a measurable function-preserving axis within the post-grokking family.

# Population, schedules, and alignment

## Population summary

We attempted 232 training runs across the schedule grid (§3.3) and the 10-seed population (§3.4). 217 runs reached the grokking threshold of test accuracy $\geq 0.95$ within their respective step budgets; 15 did not. The non-grokked runs cluster in the low-$\lambda$, low-$\eta$ corner of the schedule grid. Main-text analyses use this 217-model grokked population except where explicitly stated; §6 reports raw participation-ratio statistics on the original 117-model batch, as explained in §A.5.

### Seed list

We use 10 random seeds, divided into two batches:

- **Original batch (5 seeds):** $\{7, 13, 21, 42, 99\}$.
- **Extension batch (5 seeds):** $\{101, 137, 211, 313, 401\}$.

The extension batch was added to test the population-level findings for stability under additional initialization sampling.

### Constant weight-decay schedule grid

Constant-$\lambda$ runs use a $6 \times 3$ grid:

$\eta$ $\lambda$ 0.3 0.5 1.0 1.5 2.0 3.0 ——————- —— —— —— —— —— ——

0.01
0.02

18 configurations per seed $\times$ 10 seeds = 180 attempted constant-schedule runs.

### Cyclic weight-decay schedule grid

Cyclic-$\lambda$ runs oscillate sinusoidally between $\lambda_{\min}$ and $\lambda_{\max}$ at a fixed period:

| Range | Period (steps) | $\eta$ |
|---|---|---|
| $[0.3, 2.5]$ | 1,500 | 0.01 |
| $[0.3, 2.5]$ | 3,000 | 0.01 |
| $[0.3, 2.5]$ | 6,000 | 0.01 |
| $[0.5, 2.0]$ | 3,000 | 0.01 |

Original-batch seeds use 6 cyclic configurations each (3 periods $\times$ 2 ranges); extension-batch seeds use 4 cyclic configurations each. Seed 13 missed all original-batch cyclic configurations. Total: $5 \times 6 + 5 \times 4 - 6 = 44$ attempted cyclic-schedule runs.

### Per-batch grokking counts

The frozen population comprises 232 attempted runs split across two batches with slightly different cyclic schedule grids (§A.4). All runs are filtered by the same grokking criterion.

| Batch | Const attempted | Cyc attempted | Total | Grokked | Non-grokked |
|---|---|---|---|---|---|
| Original | 98 | 24 | 122 | 117 | 5 |
| Extension | 90 | 20 | 110 | 100 | 10 |
| **Total** | **188** | **44** | **232** | **217** | **15** |

Seed lists are given in §A.2.

Per-seed grokking counts (original batch):

@ >

p(- 8) * 0.2000 >

p(- 8) * 0.2000 >

p(- 8) * 0.2000 >

p(- 8) * 0.2000 >

p(- 8) * 0.2000@

Seed

&

Const attempted

&

Cyc attempted

&

Total attempted

&

Grokked

& 18 & 6 & 24 & 23
13 & 18 & 0 & 18 & 17
21 & 18 & 6 & 24 & 23
42 & 26 & 6 & 32 & 31
99 & 18 & 6 & 24 & 23

The five original-batch non-grokked runs comprise one per seed, all at $\lambda = 0.3, \eta = 0.005$ (constant schedule) — a single low-$\lambda$, low-$\eta$ corner of the schedule grid that fails reliably across initializations. Per-run metadata is released with the code.

**Logit-norm heterogeneity and §6 scope.** The extension batch's grokked models exhibit a substantially broader distribution of readout magnitudes than the original batch. Logit L2 norm across the 117 original-batch grokked models ranges from 554 to 1,593 (median 1,059); across the 100 extension-batch grokked models it ranges from 554 to 50,613 (median 1,505), with 10 models exceeding 5× the population median. All 10 outliers occur at $\lambda_{\min} = 0.3$, the lowest weight-decay value in the extension schedule grid, with the most extreme outlier (logit norm 50,613) coming from seed 137 at $\lambda = 0.3, \eta = 0.02$. Raw-logit participation ratio is not scale-invariant; under L2-per-model normalization on the full $n = 217$ population, the dimension ladder becomes logit PR $\approx$ 7.3, hidden PR $\approx$ 179, parameter PR $\approx$ 193. We report §6 on the $n = 117$ original-batch population to maintain interpretability of the raw-logit measurement; §§7–9 use the full 217-model population because their analyses are evaluated through held-out geometry, shuffled nulls, dimensionality controls, explicit intervention comparisons, or normalized perturbation directions rather than raw covariance PR on unnormalized logits.

### Activation-Hungarian alignment procedure

For each pair (reference, target) of grokked models we align target hidden units to reference hidden units by maximum cosine similarity over the full input grid, then propagate the permutation through fc1 output rows and fc2 input columns.

```
Input:  reference model M_ref, target model M_tgt
        full input grid X = {0, ..., p-1}^2

1.  H_ref = ReLU(fc1_ref(embed_ref(X)))           # shape (P^2, h)
2.  H_tgt = ReLU(fc1_tgt(embed_tgt(X)))           # shape (P^2, h)
3.  Normalize each column of H_ref and H_tgt to unit L2 norm.
4.  C = H_ref^T @ H_tgt                           # shape (h, h), cosine sim
5.  pi = Hungarian(-C)                            # maximize total similarity
6.  Apply pi to:
        target fc1 output weights and biases (rows)
        target fc2 input weights (columns)
        target hidden activations (output dim)
    Embedding vectors are NOT permuted (they precede the hidden layer).

Output: aligned target model M_tgt^aligned
```

The reference model is fixed before all alignment analyses (model 0; see §3.5). Cross-population alignment is performed pairwise against this single reference, not jointly.

## Local mode-connectivity endpoint pairs

The local mode-connectivity analysis in §4 measures cross-entropy loss along two interpolation paths between five same-seed grokked endpoint pairs: a straight-line linear path and a quadratic Bezier path through an optimized midpoint. Endpoint pairs are selected by maximum function-space L2 distance within each seed. This appendix lists the endpoint configurations and per-pair barrier measurements.

### Endpoint pair configurations

@ >

p(- 20) * 0.0769 >

p(- 20) * 0.0769 >

p(- 20) * 0.0769 >

p(- 20) * 0.0897 >

p(- 20) * 0.0897 >

p(- 20) * 0.0897 >

p(- 20) * 0.0897 >

p(- 20) * 0.1026 >

p(- 20) * 0.1026 >

p(- 20) * 0.1026 >

p(- 20) * 0.1026@

Seed

&

$\lambda_a$

&

$\lambda_b$

&

$\eta_a$

&

$\eta_b$

&

$\text{sched}_a$

&

$\text{sched}_b$

&

$\text{acc}_a$

&

$\text{acc}_b$

&

$\text{CE}_a$

&

$\text{CE}_b$

& 0.30 & 1.00 & 0.010 & 0.010 & const & const & 0.9955 & 1.0000 & 0.0308 & 0.0092
13 & 0.30 & 3.00 & 0.020 & 0.020 & const & const & 1.0000 & 1.0000 & 0.0030 & 0.1922
21 & 0.50 & 1.50 & 0.020 & 0.010 & const & cyc & 0.9992 & 1.0000 & 0.0040 & 0.0003
42 & 1.00 & 1.65 & 0.005 & 0.010 & const & cyc & 1.0000 & 1.0000 & 0.0117 & 0.0002
99 & 0.50 & 1.50 & 0.020 & 0.005 & const & const & 1.0000 & 1.0000 & 0.0016 & 0.0198

Both endpoints in each pair are drawn from the grokked population (acc $\geq 0.95$) and share the same initialization seed. Endpoint CE values are reported on the full 1,326-pair test set. Function-space L2 distances for the selected pairs range from 1,037 (seed 42) to 1,491 (seed 21), confirming that

endpoints are well-separated within their respective seed populations. For cyclic schedules ("cyc"), $\lambda$ is the per-run identifier; full schedule specification is in §3.3.

### Barrier measurements

@ >

p(- 8) * 0.0714 >

p(- 8) * 0.1786 >

p(- 8) * 0.2857 >

p(- 8) * 0.1786 >

p(- 8) * 0.2857@

Seed

&

linear max CE

&

linear barrier height

&

Bezier max CE

&

Bezier barrier height

& 2.540 & 2.509 & 0.073 & 0.042
13 & 2.987 & 2.795 & 0.192 & 0.000
21 & 2.604 & 2.600 & 0.087 & 0.083
42 & 2.837 & 2.826 & 0.089 & 0.078
99 & 2.801 & 2.782 & 0.089 & 0.069
**mean** & **2.754** & **2.702** & **0.106** & **0.054**

*Barrier height* is the maximum CE along the path minus the larger of the two endpoint CE values: height $= \max_\alpha \mathrm{CE}(\theta(\alpha)) - \max(\mathrm{CE}_a, \mathrm{CE}_b)$.

Across all five pairs, the linear-interpolation midpoint reaches a cross-entropy of 2.5–3.0 nats — comparable in magnitude to the cross-entropy of a uniform 47-class predictor ($\log 47 \approx 3.85$ nats) — while the endpoints themselves have CE typically below 0.03. A single Bezier midpoint optimization reduces the barrier height to below 0.1 in four of five cases and to 0.000 in the fifth.

The seed-13 result deserves a note: its Bezier *max CE* equals endpoint $b$'s CE (0.1922) to four decimal places, meaning the Bezier path nowhere exceeds the worse endpoint. The barrier height is exactly 0.000 — the path is fully sub-endpoint along its entire length.

### Bezier midpoint optimization

The midpoint $\theta_m$ of the Bezier path is found by gradient descent on a smooth approximation of the path's maximum-loss objective:

$$\theta_m^* = \arg\min_{\theta_m} \ \tau^{-1} \log \sum_{\alpha \in \mathcal{A}} \exp(\tau \cdot \mathrm{CE}(\theta(\alpha; \theta_a, \theta_m, \theta_b))) \,,$$

with endpoints $\theta_a, \theta_b$ held fixed. The path is the standard quadratic Bezier curve $\theta(\alpha) = (1-\alpha)^2\theta_a + 2\alpha(1-\alpha)\theta_m + \alpha^2\theta_b$, evaluated on a grid $\mathcal{A}$ of 21 evenly spaced $\alpha$ values from 0 to 1. The objective is a temperature-$\tau$ soft-maximum over $\mathcal{A}$.

**Hyperparameters used in our experiments:**

- Optimizer: Adam, learning rate $10^{-3}$
- Number of optimization steps: 500
- Soft-max temperature: $\tau = 10$
- Midpoint initialization: $\theta_m^{(0)} = (\theta_a + \theta_b)/2$ (the linear midpoint)

The midpoint is *not* a separately trained grokked model. It is a low-loss waypoint between two grokked endpoints, constructed only to demonstrate that the endpoints are not separated by a closed high-loss surface in their joint neighborhood (§4).

Same-seed endpoint pairs do not require unit-permutation alignment (§3.5) because they share initialization; the hidden-unit ordering is preserved throughout training within a single seed.

### Endpoint selection criterion

Within each seed's grokked population, the endpoint pair $(a, b)$ is selected to maximize L2 distance in logit-function space, where the logit-function vector for a model is its concatenated logit output across all 1,326 held-out test pairs. This criterion selects the most dispersed same-seed pair available, providing the strongest test for §4's claim: if even the maximally separated grokked endpoints lie on a low-loss curve, less-separated endpoints do too.

Bansal, Yamini, Preetum Nakkiran, and Boaz Barak. 2021. "Revisiting Model Stitching to Compare Neural Representations." In *Advances in Neural Information Processing Systems (NeurIPS)*.

Chughtai, Bilal, Lawrence Chan, and Neel Nanda. 2023. "A Toy Model of Universality: Reverse Engineering How Networks Learn Group Operations." In *International Conference on Machine Learning (ICML)*.

Conmy, Arthur, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. "Towards Automated Circuit Discovery for Mechanistic Interpretability." In *Advances in Neural Information Processing Systems (NeurIPS)*.

Draxler, Felix, Kambis Veschgini, Manfred Salmhofer, and Fred A. Hamprecht. 2018. "Essentially No Barriers in Neural Network Energy Landscape." In *International Conference on Machine Learning (ICML)*.

Elhage, Nelson, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, et al. 2021. "A Mathematical Framework for Transformer Circuits." *Transformer Circuits Thread, Anthropic*. https://transformer-circuits.pub/2021/framework/index.html.

Frankle, Jonathan, Gintare Karolina Dziugaite, Daniel M. Roy, and Michael Carbin. 2020. "Linear Mode Connectivity and the Lottery Ticket Hypothesis." In *International Conference on Machine Learning (ICML)*.

Garipov, Timur, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P. Vetrov, and Andrew Gordon Wilson. 2018. "Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs." In *Advances in Neural Information Processing Systems (NeurIPS)*.

Gromov, Andrey. 2023. "Grokking Modular Arithmetic." *arXiv Preprint arXiv:2301.02679*.

Juneja, Jeevesh, Rachit Bansal, Kyunghyun Cho, João Sedoc, and Naomi Saphra. 2023. "Linear Connectivity Reveals Generalization Strategies." In *International Conference on Learning Representations (ICLR)*.

Kornblith, Simon, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. "Similarity of Neural Network Representations Revisited." In *International Conference on Machine Learning (ICML)*.

Lenc, Karel, and Andrea Vedaldi. 2015. "Understanding Image Representations by Measuring Their Equivariance and Equivalence." In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liu, Ziming, Ouail Kitouni, Niklas Nolte, Eric Michaud, Max Tegmark, and Mike Williams. 2022. "Towards Understanding Grokking: An Effective Theory of Representation Learning." In *Advances in Neural Information Processing Systems (NeurIPS)*.

Liu, Ziming, Eric J. Michaud, and Max Tegmark. 2023. "Omnigrok: Grokking Beyond Algorithmic Data." In *International Conference on Learning Representations (ICLR)*.

Nanda, Neel, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. "Progress Measures for Grokking via Mechanistic Interpretability." In *International Conference on Learning Representations (ICLR)*.

Notsawo, Pascal Junior Tikeng, Hattie Zhou, Mohammad Pezeshki, Irina Rish, and Guillaume Dumas. 2023. "Predicting Grokking Long Before It Happens: A Look into the Loss Landscape of Models Which Grok." *arXiv Preprint arXiv:2306.13253*.

Olah, Chris, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. "Zoom in: An Introduction to Circuits." *Distill*. https://distill.pub/2020/circuits/zoom-in/.

Olsson, Catherine, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, et al. 2022. "In-Context Learning and Induction Heads." *Transformer Circuits Thread, Anthropic*. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Power, Alethea, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. 2022. "Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets." *arXiv Preprint arXiv:2201.02177*.

Rubin, Noa, Inbar Seroussi, and Zohar Ringel. 2024. "Grokking as a First Order Phase Transition in Two Layer Networks." In *International Conference on Learning Representations (ICLR)*.

Thilak, Vimal, Etai Littwin, Shuangfei Zhai, Omid Saremi, Roni Paiss, and Joshua Susskind. 2022. "The Slingshot Mechanism: An Empirical Study of Adaptive Optimizers and the Grokking Phenomenon." *arXiv Preprint arXiv:2206.04817*.

Varma, Vikrant, Rohin Shah, Zachary Kenton, János Kramár, and Ramana Kumar. 2023. "Explaining Grokking Through Circuit Efficiency." *arXiv Preprint arXiv:2309.02390*.

Zhong, Ziqian, Ziming Liu, Max Tegmark, and Jacob Andreas. 2023. "The Clock and the Pizza: Two Stories in Mechanistic Explanation of Neural Networks." In *Advances in Neural Information Processing Systems (NeurIPS)*.

Žunkovič, Bojan, and Enej Ilievski. 2022. "Grokking Phase Transitions in Learning Local Rules with Gradient Descent." *arXiv Preprint arXiv:2210.15435*.