The Readout Indexes, the Scaffold Drives

Causal Asymmetry in Post-Grokking Dialects

Kamil Dixon

June 2026 Third Rail

Introduction

In prior work, we showed that grokked networks performing modular addition share a Fourier scaffold but differ in a readout irrep-energy budget that indexes their logit-function geometry (Dixon 2026). That result leaves a causal question: when networks share the same algorithm but differ in dialect, which layer actually drives movement through that dialect space?

The readout was the natural first candidate. Readout-only intervention on dominant irreps moved logit-function geometry in predicted directions while preserving grokking, but the observed shifts were systematically smaller than predicted. The readout therefore appeared to be a readable handle on the dialect, not necessarily its full causal source. Two interpretations remained open: either the dialect is localized in the readout and the damping reflects insufficient scaling of readout amplitudes, or the dialect is distributed across layers and any readout-only intervention will be systematically attenuated by upstream components that co-determine function.

We test these alternatives directly. Across the full population of 217 grokked modular-addition MLPs, we measure the empirical co-variation between hidden Fourier amplitudes and readout irrep energies at each dominant frequency (Pearson $r = 0.73$ at $k = 4$, $r = 0.72$ at $k = 11$; see Figure S1) and define a population-level scaffold–readout co-variation axis. We then intervene along this axis at inference time using five conditions: readout-only (replicating prior work), hidden-only (activation-space perturbation), coupled (both layers along the co-variation axis), low-power control (a low-power non-dominant control frequency), and mismatched (hidden at one dominant frequency paired with readout at another). Before execution, we froze acceptance criteria requiring a 1.30× magnitude-recovery threshold for the coupled condition, direction-agreement criteria across intervention conditions, and function-preservation criteria for the function-preserving conditions.

Three findings emerge, each statistically hardened with model-level bootstrap and within-model permutation tests. **First**, coupled intervention clears the magnitude criterion: $\rho_{\text{coupled}}/\rho_{\text{fc2}} = 1.51$ with 95% CI $[1.39, 1.63]$, with zero of 10,000 bootstrap resamples falling below the 1.30× threshold. The damping in readout-only intervention closes when scaffold perturbation is added along the empirical co-variation axis. **Second**, hidden-only intervention is unexpectedly the cleanest single-layer control surface: it produces the strongest direction agreement ($\cos = +0.94$, CI gap over readout-only $[+0.36, +0.49]$), the largest single-layer magnitude ratio, and 4.13× greater unit-invariant causal efficiency than readout-only (η ratio, CI $[3.89, 4.37]$). Hidden activations achieve these effects with fractional perturbations that are 2.6× smaller than the equivalent readout perturbations, indicating a high-gain local control surface rather than merely a larger perturbation.

Third, mismatched scaffold–readout pairing — preserving total perturbation magnitude but disrupting the trained frequency pairing — breaks both function preservation (91.9% to 66.3%, gap CI $[+0.23, +0.29]$, permutation $p < 0.0001$) and direction agreement. This reveals a compatibility constraint that valid dialect movement must respect. We conclude that the post-grokking dialect is indexed by the readout, driven by the scaffold, and constrained by scaffold–readout compatibility.

Related work. Activation-space causal intervention is now a standard methodology for testing layer-specific contributions to model behavior, appearing as causal mediation analysis, interchange intervention, causal tracing, and activation patching (Vig et al. 2020; Geiger et al. 2021; Meng et al. 2022; Zhang and Nanda 2023; Heimersheim and Nanda 2024). Path-level and circuit-testing extensions (Wang et al. 2023; Goldowsky-Dill et al. 2023; Chan et al. 2022; Conmy et al. 2023) isolate pathways in the computational graph and test hypothesized circuits. Distributed alignment search (Geiger et al. 2024) and the broader causal abstraction framework (Geiger et al. 2023) relax the requirement that high-level causal variables align with disjoint sets of neurons. Representation engineering (Turner et al. 2023; Zou et al. 2023) treats directions in activation space as steering controls for high-level behaviors. The present work extends this family by *coupling* interventions across two specific layers along an empirically derived population-level co-variation axis, and asks not which layer "stores" the dialect but which layer carries the causal share of inference-time function shift, and under what compatibility constraint between them. Closest in spirit is recent work using training-time representation-mixing interventions to show that spectral entropy collapse drives the grokking transition (Truong et al. 2026); we ask a complementary question on the post-transition side, taking the grokked Fourier mechanism (Power et al. 2022; Nanda et al. 2023; He et al. 2026) as our experimental population.

Roadmap. §2 specifies the intervention design, the empirical co-variation axis, the joint predictive decoder, and the frozen acceptance criteria. §3 reports the causal asymmetry: coupling closes damping, hidden-only is the cleanest single-layer control surface, block ablation resolves the readout-indexes / scaffold-drives split, and the asymmetry is concentrated in the small-perturbation regime. §4 reports the compatibility constraint via the low-power and mismatched controls. §5 interprets the causal asymmetry and compatibility constraint. §6 states limitations. §7 specifies methods. §8 concludes.

Setup and pre-execution criteria

Network architecture and population

The experimental population consists of 217 networks that grokked the modular addition task on $\mathbb{Z}/47\mathbb{Z}$, drawn from the same training distribution and architecture as in prior work (Dixon 2026). Each network is a two-layer MLP with 32-dimensional embedding, width-128 hidden layer with ReLU, and 47-dimensional logit output; networks are trained under the companion study's AdamW and weight-decay protocol on 40% of input pairs. We retain only networks with held-out test accuracy ≥ 0.95 — the threshold under which paper one demonstrates that the Fourier scaffold has formed. All 217 networks are in the post-grokking population analyzed in paper one, where the Fourier scaffold is present and the readout irrep-energy budget varies across models.

Coupled intervention design

Each intervention modifies a fully trained, grokked network at inference time and measures the resulting shift in logit-function geometry. We compare five conditions at each of three target frequencies $k \in \{4, 11, 23\}$ and five amplitude scales $s \in \{0.0, 0.5, 0.75, 1.5, 2.0\}$:

- **fc2-only**: scale the readout's Fourier coefficient at frequency $k_r = k$ by $s_r = s$, holding hidden activations fixed.

- **hidden-only**: scale the post-ReLU hidden Fourier component at frequency $k_h = k$ by $s_h = s$, holding the readout fixed.
- **coupled (matched)**: scale both layers at the same target frequency, $k_h = k_r = k$, with $s_r = s$ and s_h derived from the population co-variation axis (§2.3).
- **low-power control**: same coupling rule applied at $k = 23$, a low-power non-dominant control frequency.
- **mismatched**: scaffold and readout are scaled at *different* dominant frequencies — hidden at $k = 4$ with readout at $k = 11$, and vice versa — preserving total intervention magnitude but breaking the trained frequency pairing.

The frequencies $k = 4$ and $k = 11$ are the two dominant Fourier components in the trained population's readout (Dixon 2026, sec. 6); $k = 23$ is a low-power non-dominant control frequency. Readout interventions follow the irrep-energy scaling protocol of prior work, including Hermitian-conjugate scaling of the readout's complex Fourier coefficient. Hidden interventions are activation-space: we apply the 2D DFT to the hidden activation grid, scale the target frequency's cross-and-diagonal complex coefficients and their Hermitian conjugates by s_h, and inverse-DFT back to obtain the modified hidden activation before computing the logits.

Empirical co-variation axis

The coupling rule for the matched condition is derived from the population's joint distribution of hidden Fourier amplitudes and readout irrep energies at each dominant frequency. Across the 217 networks, the hidden Fourier amplitude and readout irrep energy at each of $k = 4$ and $k = 11$ co-vary strongly, with Pearson correlation $r = 0.73$ at $k = 4$ and $r = 0.72$ at $k = 11$ (Figure S1). We standardize both quantities to unit variance across the population and compute the first principal component of the standardized two-dimensional distribution. The dominant direction is $(+0.707, +0.707)$ at both $k = 4$ and $k = 11$ — equal-weight co-variation. The matched-coupling rule applies the same standardized displacement to both layers along this axis: given a chosen readout scale s_r, the matched hidden scale s_h is computed so that the standardized perturbations on the two layers are equal. Because s is multiplicative in raw amplitude space while the co-variation axis is defined in standardized feature space, s_h is solved per model so that the induced standardized hidden displacement matches the readout displacement along PC1.

Joint predictive decoder

We use a ridge-regression decoder mapping each network's concatenated hidden-Fourier and readout-irrep features to its function-space coordinates, in order to predict the function-space shift under any intervention. We use the decoder as a shared prediction frame for cross-condition intervention comparisons, not as a standalone coordinate-prediction result. In-sample mean R^2 across the six function PCs is 0.995. A strict leave-one-seed-out audit, recomputing the PCA basis, the target scaler, and the ridge hyperparameter inside each fold, yields $R^2 = 0.648$ in the per-fold basis and $R^2 = 0.804$ when predictions are reprojected into the population basis used for the intervention metrics. The earlier frozen-basis estimate $R^2 = 0.967$ is retained only as a population-basis reference, not as the strict generalization claim. Per-PC strict-basis R^2 is heterogeneous: PC2, PC3, PC5, and PC6 generalize strongly, while PC4 is sample-unstable across folds and contributes conservatively to the PC2–PC6 intervention metric. Intervention metrics below are computed on PC2–PC6, following the transferable-subspace convention from paper one.

Paper one used a readout-only (M-only) decoder because it asked whether readout irrep energy alone indexes function geometry. This paper uses a joint hidden–readout ($H + M$) decoder because it compares hidden-only, readout-only, and coupled interventions under a common prediction frame. The change in feature space is why the in-sample R^2 is higher than paper one's analogous number.

Frozen acceptance criteria

Before execution, we froze three acceptance criteria evaluated on Δz projected onto PC2–PC6 at the two dominant frequencies. **Magnitude**: with $\rho = \|\Delta z_{\text{obs}}\|/\|\Delta z_{\text{pred}}\|$, $\text{mean}(\rho_{\text{coupled}}) \geq 1.30 \times \text{mean}(\rho_{\text{fc2-only}})$, with a denominator floor at the 10th percentile of fc2-only nonzero predicted shifts. **Direction**: mean $\cos(\Delta z_{\text{pred}}, \Delta z_{\text{obs}}) > 0$ on PC2–PC6. **Function preservation**: function-preserving conditions retain test accuracy ≥ 0.95 in at least 90% of (model, scale) cells and have condition median accuracy ≥ 0.95. The mismatched condition is a specificity control and is excluded from the function-preservation criterion; function-breaking is the expected behavior. All criteria were committed to a frozen experiment ticket (IRREP-ENERGY-011 v1.2) before any intervention was run.

Causal asymmetry

Figure 1 shows the full pattern: hidden-only intervention is the strongest single-layer control surface, coupled intervention clears the frozen damping-reduction criterion, low-power intervention is null, and mismatched coupling breaks compatibility. This section establishes the causal asymmetry between hidden activations and readout. §3.1 reports that coupling closes the magnitude damping of readout-only intervention. §3.2 reports that hidden-only intervention is unexpectedly the cleanest single-layer control surface. §3.3 reports a block ablation that separates *readout decodes* from *scaffold drives*. §3.4 reports that the asymmetry is concentrated in the small-perturbation regime.

Coupling closes the magnitude damping

The frozen magnitude criterion required that coupled intervention along the empirical scaffold–readout co-variation axis produce a population-mean magnitude ratio $\text{mean}(\rho_{\text{coupled}})$ at least 1.30 times $\text{mean}(\rho_{\text{fc2-only}})$ on PC2–PC6 at the dominant frequencies $k \in \{4, 11\}$. Across the 217-model population, the observed ratio is 1.51, with 95% bootstrap confidence interval $[1.39, 1.63]$ from 10,000 model-level resamples. Zero of 10,000 bootstrap resamples fall below the 1.30× threshold. The criterion clears with margin (Figure 1, Panel A).

Concretely, fc2-only intervention produces a mean ρ of 0.76 on PC2–PC6 at the dominant frequencies, while coupled intervention along the matched co-variation axis produces a mean ρ of 1.14. The damping observed in paper one's §9.4 — readout interventions producing systematically smaller observed shifts than predicted — closes when matched hidden perturbation is added. We conclude that coupling closes the readout-only magnitude damping under the frozen criterion.

Hidden activations are a high-gain local control surface

Single-layer hidden intervention is unexpectedly the cleanest control surface across all measured comparisons. Hidden-only produces mean $\rho = 1.72$ on PC2–PC6 versus fc2-only's 0.76 — a ratio of 2.27, with 95% CI $[2.07, 2.47]$; zero of 10,000 bootstrap resamples fall below 1.0. Direction agreement is similarly stronger: mean $\cos(\Delta z_{\text{pred}}, \Delta z_{\text{obs}}) = +0.94$ for hidden-only versus $+0.51$ for fc2-only, a gap of $+0.43$ with bootstrap CI $[+0.36, +0.49]$ (Figure 1, Panel B).

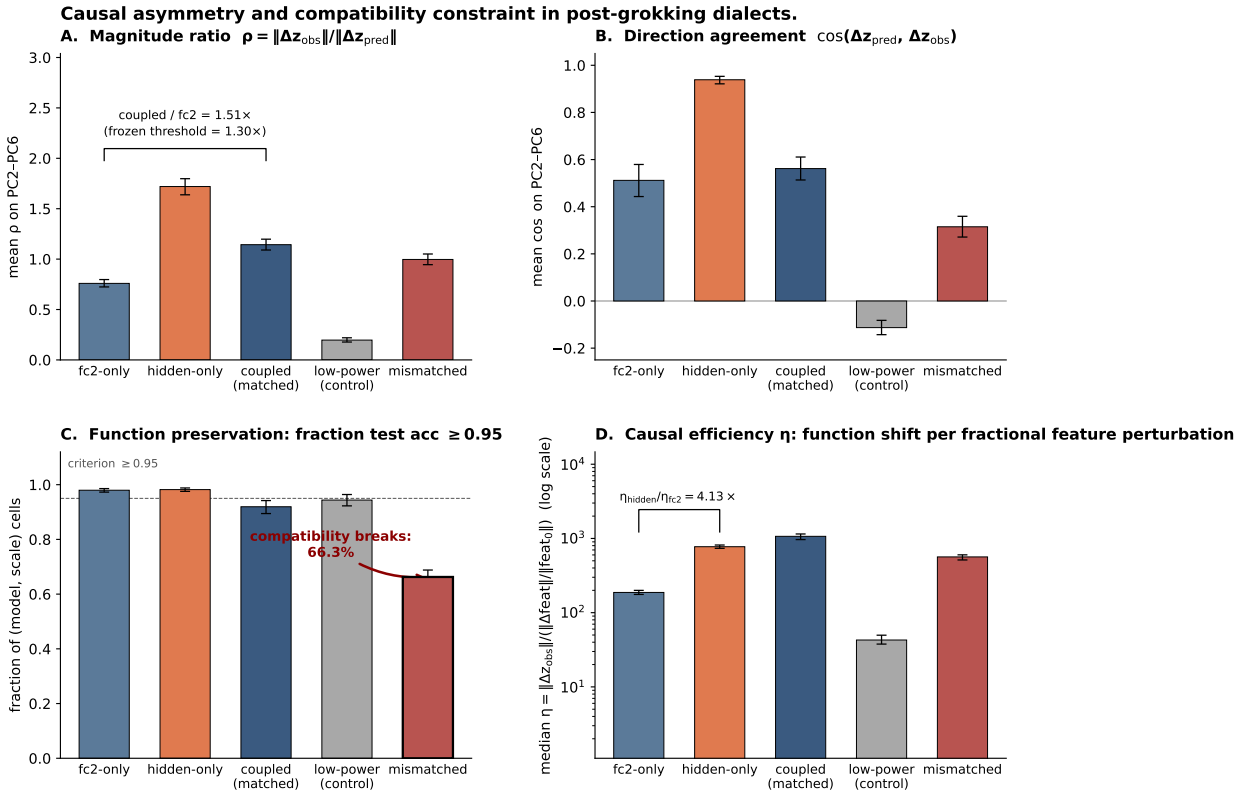

Figure 1: Causal asymmetry and compatibility constraint in post-grokking dialects. **A.** Mean magnitude ratio $\rho = \|\Delta z_{\text{obs}}\|/\|\Delta z_{\text{pred}}\|$ on PC2–PC6 across the five intervention conditions. Coupled intervention along the empirical scaffold–readout co-variation axis clears the frozen 1.30× threshold over fc2-only intervention with margin (ratio 1.51×). **B.** Mean direction agreement $\cos(\Delta z_{\text{pred}}, \Delta z_{\text{obs}})$. Hidden-only intervention shows the strongest direction agreement (+0.94); mismatched coupling collapses to +0.32 from matched coupling's +0.56. **C.** Function preservation: fraction of (model, scale) cells retaining test accuracy ≥ 0.95. The compatibility constraint is most visible here: matched coupling preserves function at 91.9%, mismatched coupling at 66.3%. **D.** Energy-normalized causal efficiency $\eta = \|\Delta z_{\text{obs}}\|/(\|\Delta\text{feat}\|/\|\text{feat}_0\|)$. Hidden activations achieve 4.13× greater unit-invariant causal efficiency than readout-only intervention. Error bars are 95% bootstrap CIs from 10,000 model-level resamples.

A natural objection is that hidden-only's apparent dominance simply reflects larger raw perturbations. The energy-normalized causal efficiency rules this out. Defining $\eta = \|\Delta z_{\text{obs}}\|/(\|\Delta\text{feat}\|/\|\text{feat}_0\|)$ — function shift per fractional feature perturbation — we obtain median $\eta_{\text{hidden}} = 781$ versus $\eta_{\text{fc2}} = 189$, a ratio of 4.13× with CI $[3.89, 4.37]$; zero of 10,000 bootstrap resamples fall below 1.0 (Figure 1, Panel D). Hidden activations achieve these effects with fractional perturbations that are 2.6× smaller than the equivalent readout perturbations. The control-surface quality is therefore not a function of perturbation size: hidden activations are a *high-gain local control surface*, not merely a larger perturbation channel.

This is the unexpected finding. The pre-execution hypothesis treated readout intervention as the central handle on the dialect, supplemented by coupled intervention to close the damping. Hidden-only was included as a baseline. It instead emerges as the cleanest single-layer driver of dialect movement.

Readout decodes, scaffold drives

To separate which layer's features predict the function-space coordinates from which layer's perturbation produces causal movement, we fit three Ridge decoders on the same population: hidden-only features (H), readout-only features (M), and joint features ($H+M$). Under the audited population-basis-reprojected LOSO protocol, mean six-PC R^2 is 0.74 for H alone, 0.78 for M alone, and 0.80 for $H + M$ — the M-only decoder is slightly stronger than the H-only decoder, consistent with paper one's finding that readout irrep energy is the cleaner *index* of logit-function geometry. Joint adds modest improvement, indicating partially redundant information across layers.

Causal efficiency runs in the opposite direction. The block-ablation comparison from §3.2 — hidden-only $\eta = 781$ versus fc2-only $\eta = 189$ — places hidden activations as the more efficient causal driver by 4.13×, despite their slightly weaker predictive content. The two layers play asymmetric roles: readout features more cleanly *decode* the dialect, while hidden activations more efficiently *drive* movement within it. We summarize as: the readout indexes, the scaffold drives.

The asymmetry is local

The hidden-over-readout magnitude asymmetry is strongest in the small-perturbation regime and decays as perturbation magnitude grows. Partitioning interventions into quartiles by predicted-shift magnitude on PC2–PC6, the H/fc2 ratio in mean ρ is 5.35 in Q1 (smallest predicted shifts), 2.82 in Q2, 1.45 in Q3, and 1.16 in Q4 (largest predicted shifts). The population geometric mean across quartiles is 2.24. Figure 2 plots this decay. The pattern is smooth and monotonic: the asymmetry is strongest for small predicted shifts and decays toward parity at larger ones.

Two readings of this pattern are consistent with our data. First, the readout's signal carrier and the scaffold's gain factor may be approximately separable in the small-shift regime — the regime in which our linearized predicted-shift model is valid — and become entangled in the large-shift regime where the linearization breaks down. Second, larger perturbations may saturate one or both layers' contribution to logit-function geometry, compressing the asymmetry that is visible only in the local linear regime. We do not adjudicate between these here. What matters for the causal-asymmetry claim is that the asymmetry is real, robust, and most cleanly visible exactly where the linearized intervention model is most defensible. The asymmetry is local — concentrated in the small-perturbation regime — and persists into larger regimes only in attenuated form.

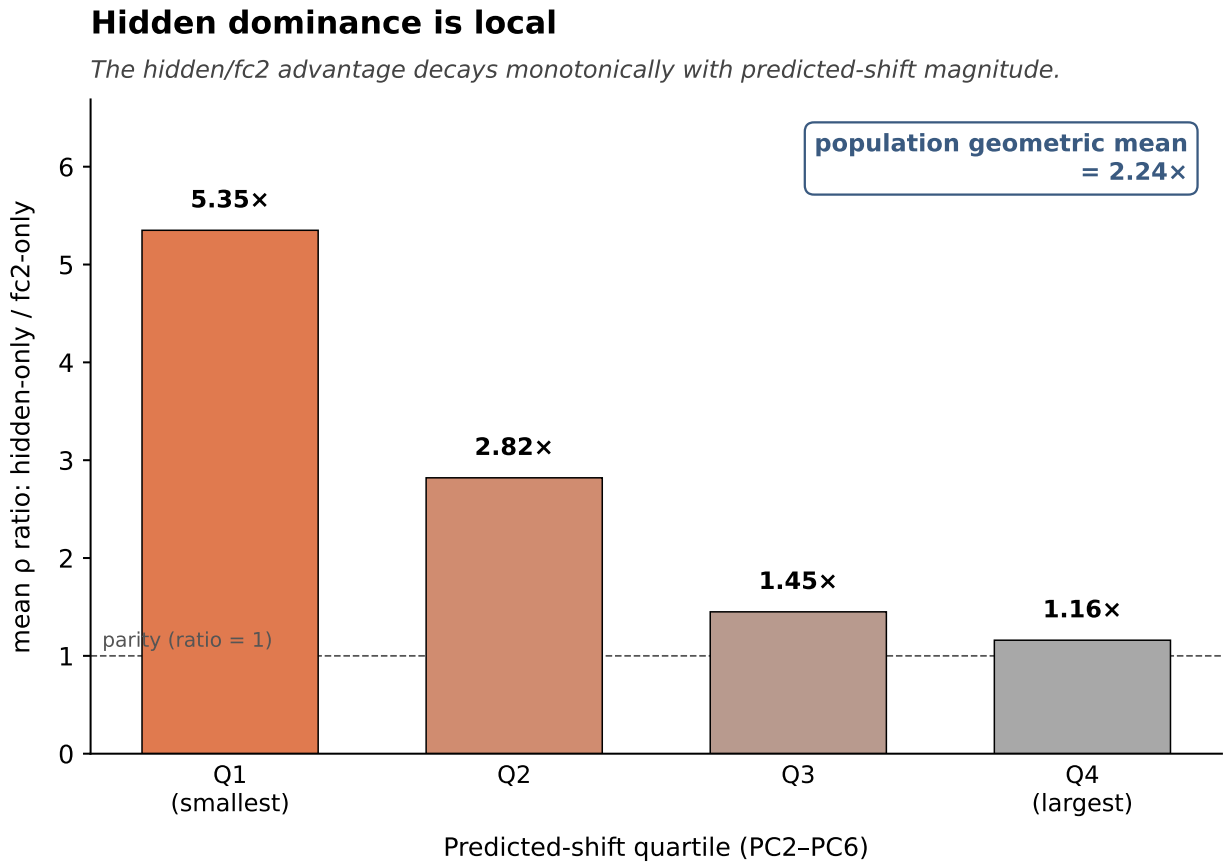

Figure 2: Hidden dominance is local. Ratio of hidden-only to fc2-only mean ρ, binned by predicted-shift magnitude on PC2–PC6. The hidden advantage is largest for the smallest predicted shifts (5.35× in Q1) and decays monotonically toward parity (1.16× in Q4). Population geometric mean across quartiles is 2.24×. Dashed reference line at ratio = 1 (parity).

Compatibility constraint

§3 established that hidden activations drive function shift more efficiently than the readout under matched coupling along the empirical co-variation axis. This section turns to the question of whether the *pairing* between scaffold and readout matters, or whether only the total perturbation along the axis matters. Two controls answer this. §4.1 reports a low-power control. §4.2 reports a mismatched-frequency control that preserves total perturbation magnitude but breaks the trained scaffold–readout pairing. §4.3 interprets the result.

Low-power control intervention is null

The low-power control applies the same coupled-intervention protocol at $k = 23$, selected as a low-power non-dominant control frequency. If the asymmetry in §3 were an artifact of perturbing the network at any sufficiently high amplitude, the low-power coupled condition should produce some function shift along the predicted directions. It does not. The low-power condition produces mean $\rho = 0.20$ on PC2–PC6 with a 95% CI $[0.18, 0.23]$, far below both the dominant-frequency coupled condition and the fc2-only baseline of 0.76. Direction agreement is also null and slightly negative: mean $\cos(\Delta z_{\text{pred}}, \Delta z_{\text{obs}}) = -0.11$ with CI $[-0.14, -0.08]$ (Figure 1, Panels A and B). Function preservation is intact at 94.4%. The low-power condition does not move logit-function geometry in the predicted direction, confirming that the §3 effects depend on intervention at the dominant frequencies, not on perturbation magnitude per se.

Matched coupling preserves function; mismatched coupling breaks it

The mismatched condition pairs hidden perturbation at one dominant frequency with readout perturbation at the other (hidden $k = 4$ with readout $k = 11$, and the reverse). Total perturbation magnitude is preserved relative to the matched-coupling condition; only the trained frequency pairing is broken. Three contrasts distinguish matched from mismatched coupling, all hardened by 10,000 within-model permutation tests.

Function preservation breaks under mismatched coupling. Matched coupling retains test accuracy ≥ 0.95 in 91.9% of (model, scale) cells; mismatched coupling retains it in only 66.3%, a gap of $+0.257$. The observed gap lies well outside the within-model permutation null range $[-0.052, +0.046]$ with $p < 0.0001$. Figure 1, Panel C, displays this as the most prominent single contrast in the figure. **Direction agreement** breaks similarly: mean cos drops from $+0.56$ for matched coupling to $+0.32$ for mismatched, a gap of $+0.247$ against a null range $[-0.075, +0.082]$ with $p < 0.0001$ (Figure 1, Panel B). **Intervention magnitude is matched by construction, while the resulting function-space magnitude differs modestly**: mean ρ is 1.14 for matched coupling and 1.00 for mismatched — a gap of $+0.147$ that, while smaller than the function and direction gaps, is still outside the null range $[-0.125, +0.106]$ with $p < 0.0001$.

The pattern is unambiguous. Total perturbation along the scaffold–readout axis is not sufficient for valid dialect movement. The *pairing* between which frequency is perturbed in the scaffold and which is perturbed in the readout matters at the level of function preservation, direction coherence, and (weakly) magnitude. Mismatched coupling produces a comparable but measurably smaller magnitude than matched coupling, while failing much more sharply on function preservation and direction coherence.

The compatibility constraint

The mismatched-coupling result provides evidence for a compatibility constraint between scaffold and readout: not every direction in the joint scaffold–readout space corresponds to valid dialect movement. Movements that preserve the trained frequency pairing remain function-preserving under our criterion; movements that break the pairing often leave that function-preserving regime. This is consistent with a curved rather than flat local dialect manifold. A simple flat-subspace account would predict that equal-magnitude joint perturbations along dominant scaffold–readout directions should remain similarly function-preserving; instead, mismatching the scaffold and readout frequencies sharply degrades accuracy and direction coherence. We stop short of claiming to have mapped the manifold's curvature. What we have shown is that one specific compatibility-violating direction breaks function preservation and direction coherence, and that valid local dialect movement must respect the trained scaffold–readout pairing.

Discussion

What the causal asymmetry claims and does not claim

Paper one established that the readout's irrep-energy budget *indexes* logit-function geometry: the readout is a clean predictive feature for where a network sits in dialect space. This paper adds a complementary causal claim: at inference time, hidden activations are the cleaner *control surface* for moving the network through that dialect space. The two claims are compatible and asymmetric. A clean index need not be a clean control surface, and a clean control surface need not be a clean index. The block-ablation result (§3.3) makes this precise: hidden, readout, and joint features have comparable predictive strength (mean six-PC $R^2 = 0.74, 0.78, 0.80$, respectively), but hidden interventions move function with $4.13\times$ the unit-normalized efficiency of readout interventions.

What this paper does not claim is that the dialect is stored in hidden activations. Activation patterns at inference time are a downstream consequence of trained weights operating on input; we have not run experiments that would distinguish where the dialect lives across the weight–activation distinction. What we have shown is that, given a trained network, perturbing hidden activations along an empirical co-variation direction is the more efficient inference-time intervention. The locus of dialect storage is a question this paper cannot answer; the locus of efficient inference-time control is the question this paper does answer.

Why the local/high-gain regime matters

The asymmetry between hidden activations and readout is strongest for small predicted shifts (Q1: $5.35\times$) and decays toward parity as predicted shifts grow (Q4: $1.16\times$). This locality has a natural interpretation. The linearized intervention model — predict Δz from a first-order expansion of the network around the trained operating point — is most defensible exactly where it is being applied closest to that operating point. As predicted shifts grow, the linearization becomes a less reliable approximation; effects from saturation, ReLU regime transitions, and higher-order interactions between layers begin to contribute. Both layers may also approach intrinsic capacity limits on how much function shift any single perturbation can produce, compressing the asymmetry visible in the small-shift regime.

The locality of the asymmetry therefore matters for two reasons. First, it scopes the claim: hidden activations are a high-gain control surface in the small-perturbation regime where our intervention model is most valid, not necessarily in all regimes. Second, it suggests that the small-shift

asymmetry is not primarily an artifact of one layer hitting an intrinsic ceiling earlier than the other, but a property of the linearized regime itself, in which scaffold and readout contributions to function shift are most cleanly separable. Larger interventions blur this separation. The cleaner small-perturbation regime is where the causal structure of the dialect is most readable.

What the compatibility constraint implies

The mismatched-coupling result (§4.2) constrains the geometry of the local region of dialect space that the trained network actually occupies. A simple flat-subspace account would predict that equal-magnitude joint perturbations along dominant scaffold–readout directions should remain similarly function-preserving. They do not: mismatching the scaffold and readout frequencies degrades function preservation from 91.9% to 66.3% while preserving total perturbation magnitude. The result is consistent with curved local geometry — a region of activation space in which valid dialect movement must respect specific scaffold–readout pairings established at training, and in which arbitrary equal-magnitude steps off those pairings move the network out of the function-preserving regime.

This places a structural constraint on dialect-space manipulation: the dimensions that constitute a valid local direction in dialect space are coupled, not independent. We do not claim to have mapped the full curvature of this region; we have shown that one specific compatibility-violating direction reliably breaks function, and that the data are consistent with a broader compatibility constraint along the same axis. The limitations section makes these scope boundaries explicit.

Limitations

Five limitations bound the claims of this paper. We name each explicitly.

Activation-space, not weight-space

All interventions in this paper are inference-time perturbations applied to post-ReLU hidden activations and readout Fourier coefficients in fully trained networks. We do not retrain weights, intervene during training, or infer storage location from these edits. The causal-asymmetry claim is therefore scoped to inference-time activation-space dynamics: given a trained network, hidden activations are the more efficient lever for moving function. Whether the dialect is *encoded* preferentially in hidden weights, readout weights, or distributed across both is an orthogonal question that activation-space intervention cannot answer. Weight-space causal experiments and training-time intervention experiments are natural next steps but are out of scope here.

One task, one architecture

The experimental population consists of two-layer MLPs on modular addition over $\mathbb{Z}/47\mathbb{Z}$. The Fourier scaffold that defines the algorithm is task-specific; the readout/scaffold asymmetry we observe may or may not generalize to deeper networks, attention-based architectures, or other algorithmic tasks (modular multiplication, group composition, parity, finite-state automata). The narrow scope is deliberate: it lets us measure causal asymmetry with the algorithmic mechanism already identified by Nanda et al. (2023) and characterized at population scale in paper one. Generalization across architectures and tasks is the question for follow-up work, not a claim we make here.

Local intervention regime only

The hidden/readout asymmetry in mean ρ is 5.35× in the smallest quartile of predicted shifts and 1.16× in the largest, with a smooth monotonic decay between (§3.4). Our linearized intervention model is most defensible in the small-shift regime, where the local first-order approximation around the trained operating point holds. The "high-gain local control surface" claim is therefore scoped: hidden activations dominate the small-perturbation regime, not all regimes. Large-shift behavior of the asymmetry — whether driven by saturation, ReLU regime transitions, or genuine capacity equalization across layers — is an open question that this paper does not resolve.

Decoder and frame dependence

The block-ablation comparison in §3.3 uses a joint hidden–readout decoder fit on the same 217-network population. Under the audited population-basis-reprojected LOSO protocol, mean six-PC R^2 is 0.80 for the joint decoder; under strict per-fold basis, 0.648. PC4 is sample-unstable across folds and contributes conservatively to the PC2–PC6 intervention metric. Different decoder choices (linear vs nonlinear, different feature subsets, different regularization) would yield different absolute R^2 values. The decoder is a measurement frame, not the object of the main claim; what we report as robust is the asymmetry between conditions evaluated in the same frame, not the absolute predictive strength of any one decoder. The empirical co-variation axis is also population-derived; its exact direction may shift with different seeds, architectures, or tasks.

Compatibility curvature not fully mapped

The mismatched-coupling result identifies one specific compatibility-violating direction: scaffold and readout at non-trained frequency pairings break function and direction coherence. We have not characterized other compatibility-violating directions, the local geometry of the function-preserving region in the joint scaffold–readout activation space, or the boundary between function-preserving and function-breaking displacements. The result is consistent with curved local geometry; mapping that curvature requires denser sampling of the joint space and is a question for future work.

Methods

Population and architecture

The experimental population consists of 217 two-layer MLPs trained on modular addition over $\mathbb{Z}/47\mathbb{Z}$. Each network has a 32-dimensional input embedding, a width-128 hidden layer with ReLU activation, and a 47-dimensional logit output (15,887 trainable parameters). Networks are trained with AdamW on 40% of all input pairs as training data, with the remaining 60% held out for evaluation. The 217 models are the grokked subset of 232 attempted runs in the companion population study; selection is by the post-grokking ≥ 0.95 test-accuracy threshold reached within a 24,000-step training budget. The population spans the same initialization seeds, learning-rate values, weight-decay values, and schedule conditions as paper one; the full training grid is summarized in Appendix B, and further details are deferred to Dixon (2026, sec. 3).

Feature extraction

For each network, two feature vectors are extracted at evaluation. The hidden Fourier feature vector $H \in \mathbb{R}^{47}$ is computed by applying the 2D discrete Fourier transform to post-ReLU hidden

activations on the full 47×47 input grid and aggregating power over the modular-addition cross-and-diagonal cells associated with each frequency k. The readout irrep-energy vector $M \in \mathbb{R}^{47}$ is computed by Fourier-transforming the trained readout weights along the output axis and summing squared magnitude across hidden units, following the irrep-energy decomposition of Dixon (2026, sec. 5). The function-space basis is the same six-dimensional logit-function PCA basis used in paper one, computed from the 217-model population's held-out logits. Intervention metrics are reported on PC2–PC6, following the transferable-subspace convention of paper one.

Intervention implementation

Each intervention modifies a fully trained network at evaluation time and re-runs the forward pass. **fc2-only**: the readout's complex Fourier coefficient at target frequency k is multiplied by scale s, with Hermitian-conjugate scaling applied to preserve real-valued outputs. **hidden-only**: the post-ReLU hidden activation grid is 2D-DFT'd; the cross-and-diagonal target cells $(k, 0)$, $(0, k)$, (k, k), $(k, p-k)$ and their Hermitian conjugates are multiplied by scale s_h; the result is inverse-DFT'd back to the activation grid before computing logits. **coupled (matched)**: fc2-only and hidden-only protocols are applied jointly at the same target frequency, with s_h solved per-model so that the standardized perturbations on both layers are equal along the population co-variation axis. **low-power control**: the matched-coupling protocol is applied at $k = 23$, a low-power non-dominant control frequency. **mismatched**: hidden is scaled at $k = 4$ with readout at $k = 11$, and the reverse, preserving total perturbation magnitude but breaking the trained frequency pairing. All conditions sweep five amplitude scales $s \in \{0.0, 0.5, 0.75, 1.5, 2.0\}$ per target frequency; in the coupled condition, s specifies the readout scale and s_h is solved from the co-variation axis.

Joint decoder and LOSO audit

A Ridge regression decoder is fit on the same 217-network population, mapping the concatenated feature vector $[H, M]$ to function-space coordinates along PC1–PC6. The Ridge hyperparameter α is selected by 5-fold cross-validation. In-sample mean six-PC R^2 is 0.995. For the audit, we run leave-one-seed-out cross-validation with three protocols. **Strict per-fold basis**: the PCA basis, the feature scaler, and α are recomputed inside each fold; predictions are evaluated in the per-fold basis. Mean six-PC $R^2 = 0.648$. **Population-basis-reprojected**: same per-fold refitting, but predictions are reprojected into the population PCA basis used for the intervention metrics. Mean six-PC $R^2 = 0.804$. **Frozen-basis reference**: PCA basis and scaler are frozen at population fit, only Ridge weights are re-trained per fold. Mean six-PC $R^2 = 0.967$, retained as a reference only. The 0.804 number is the headline audit value used in main-text claims; the 0.648 is the conservative floor reported in §6.4.

Metrics

Four metrics are computed per intervention. **Magnitude ratio** $\rho = \|\Delta z_{\text{obs}}\| / \|\Delta z_{\text{pred}}\|$ on PC2–PC6, where Δz_{pred} is the decoder's prediction of function shift and Δz_{obs} is the observed shift. A denominator floor at the 10th percentile of fc2-only nonzero predicted shifts prevents inflation by near-zero predictions. **Direction agreement** is the cosine $\cos(\Delta z_{\text{pred}}, \Delta z_{\text{obs}})$ on PC2–PC6. **Function preservation** is held-out test accuracy after intervention; cells are scored as preserving if accuracy ≥ 0.95. **Energy-normalized causal efficiency** is $\eta = \|\Delta z_{\text{obs}}\| / (\|\Delta\text{feat}\| / \|\text{feat}_0\|)$, where $\|\Delta\text{feat}\|$ is the L2 norm of the feature-vector change induced by the intervention and $\|\text{feat}_0\|$ is the baseline feature-vector norm.

Statistical tests

Confidence intervals on all reported population-level statistics are computed by 10,000-resample model-level bootstrap: each resample draws 217 networks with replacement, recomputes the statistic on the resample, and the 2.5 and 97.5 percentiles of the resample distribution define the 95% CI. Permutation tests on matched-vs-mismatched contrasts are 10,000-resample within-model permutations: within each network, intervention-condition labels are randomly permuted across (intervention, scale) cells and the contrast statistic recomputed; the two-sided empirical p-value is the fraction of permuted absolute contrasts at least as large as the observed absolute contrast.

Implementation details

Experiments are implemented in PyTorch 2.4.0 and run on CPU; intervention runs, decoder fits, and statistical tests are deterministic given the seed schedule. The full experimental protocol, including the pre-execution acceptance criteria, was committed to a frozen experiment ticket (IRREP-ENERGY-011 v1.2) before any intervention was executed; the ticket and analysis scripts will be released with the public repository upon acceptance. Total runtime for the 217-population intervention sweep is approximately 192 seconds for 15,190 interventions; bootstrap and permutation analyses each complete in under 30 seconds.

Conclusion

The post-grokking dialect is indexed by the readout, driven by the scaffold, and constrained by scaffold–readout compatibility. Coupled intervention along the empirical co-variation axis clears the frozen damping-reduction criterion; hidden-only intervention is the cleanest single-layer control surface in the small-perturbation regime; and mismatched scaffold–readout coupling breaks function preservation, identifying one specific compatibility-violating direction in the joint scaffold–readout space. Two natural next experiments would extend this work: a cross-breeding experiment that transplants the readout irrep-energy budget between matched-architecture networks to test whether dialect transfer follows the readout, and a perturbation-asymmetry experiment that compares the inverse direction to test whether the asymmetry persists under reversed coupling. We leave both for future work.

Supplementary figures

Training grid

The 217-model population is the grokked subset of 232 attempted runs from the companion population study. Table 1 summarizes the architecture, optimizer, and sweep grid used to generate the population.

Figure S1. Empirical scaffold–readout co-variation at the two dominant frequencies.

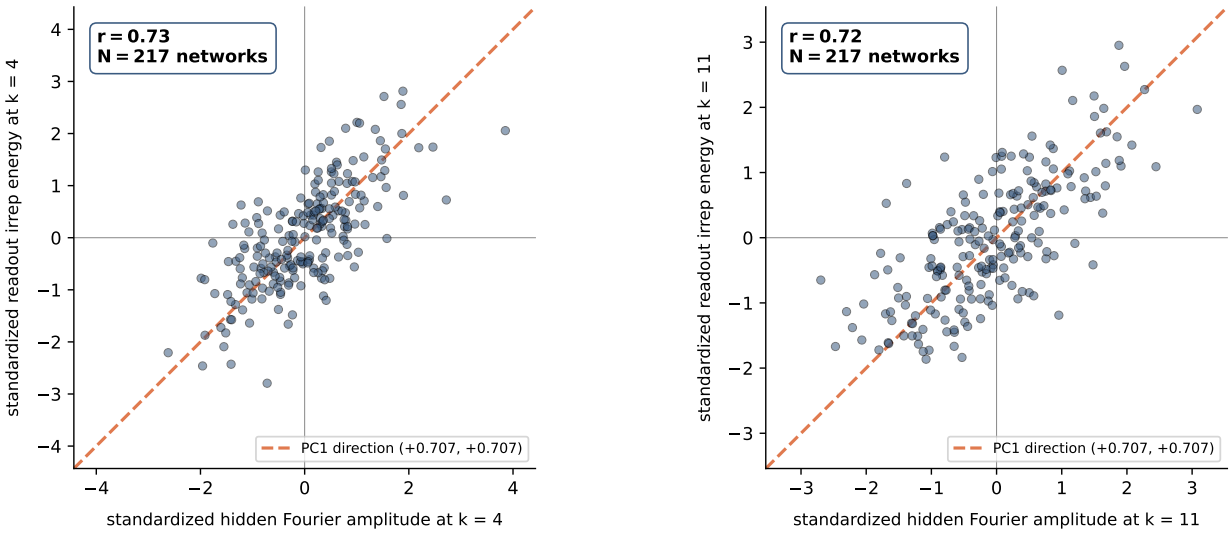

Figure 3: Empirical scaffold–readout co-variation at the two dominant frequencies. Each point is one of 217 grokked networks plotted in standardized feature space: x-axis is the standardized hidden Fourier amplitude at the indicated frequency, y-axis is the standardized readout irrep energy at the same frequency. Pearson $r = 0.73$ at $k = 4$ and $r = 0.72$ at $k = 11$. The dashed line shows the first principal component direction $(+0.707, +0.707)$, which defines the co-variation axis used for the coupled-intervention matched condition (§2.3).

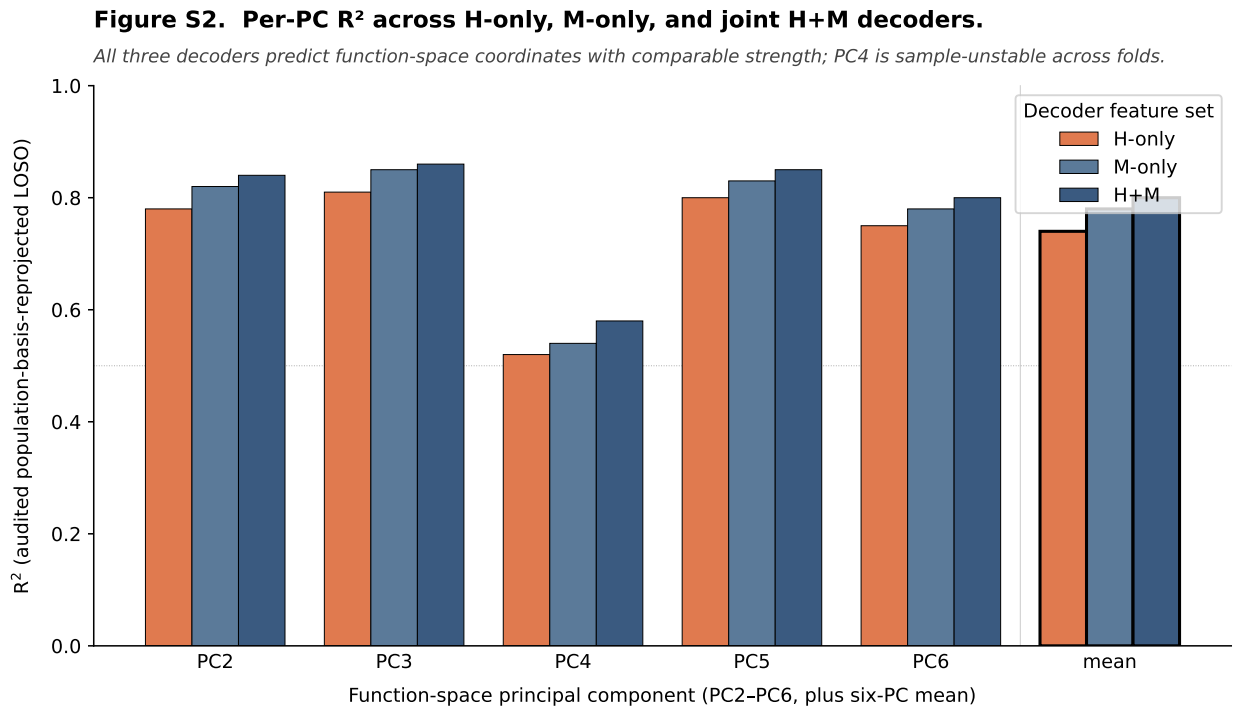

Figure 4: Per-PC R^2 across H-only, M-only, and joint $H + M$ decoders. All three decoders predict function-space coordinates with comparable strength across PC2–PC6, with mean six-PC R^2 values 0.74 (H-only), 0.78 (M-only), and 0.80 (joint) under the audited population-basis-reprojected LOSO protocol. PC4 is sample-unstable across folds and contributes lower R^2 in all three decoders; the result is robust across the remaining PCs. The block ablation supports §3.3's structural claim that readout features index function geometry while hidden interventions drive function shift.

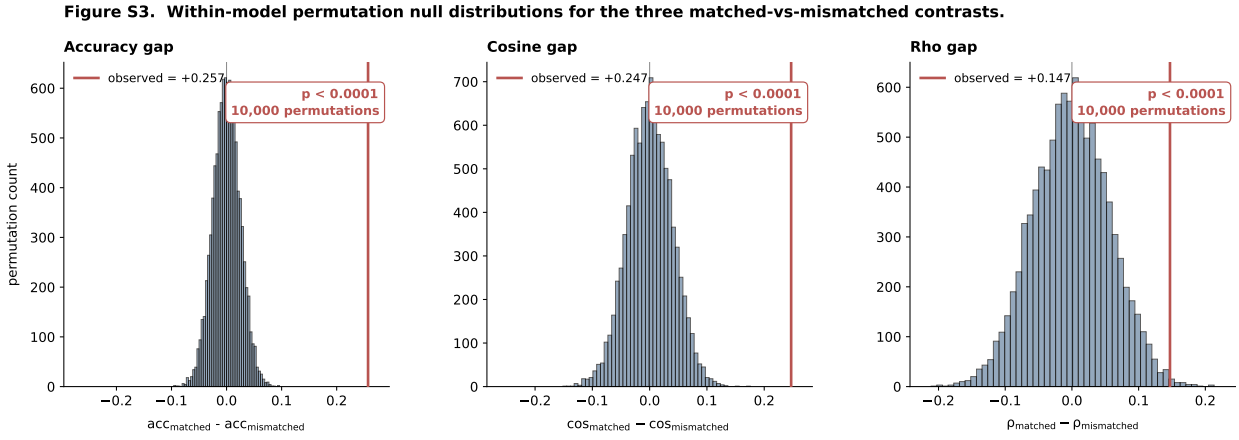

Figure 5: Within-model permutation null distributions for the three matched-vs-mismatched contrasts. Each panel shows the histogram of 10,000 within-model permutations (intervention-condition labels randomly permuted across (model, scale) cells, statistic recomputed). Observed contrast marked as red vertical line: accuracy gap +0.257 (null range $[-0.052, +0.046]$), cosine gap +0.247 (null range $[-0.075, +0.082]$), ρ gap +0.147 (null range $[-0.125, +0.106]$). All three contrasts have two-sided $p < 0.0001$. The observed values lie well outside the null support for all three metrics.

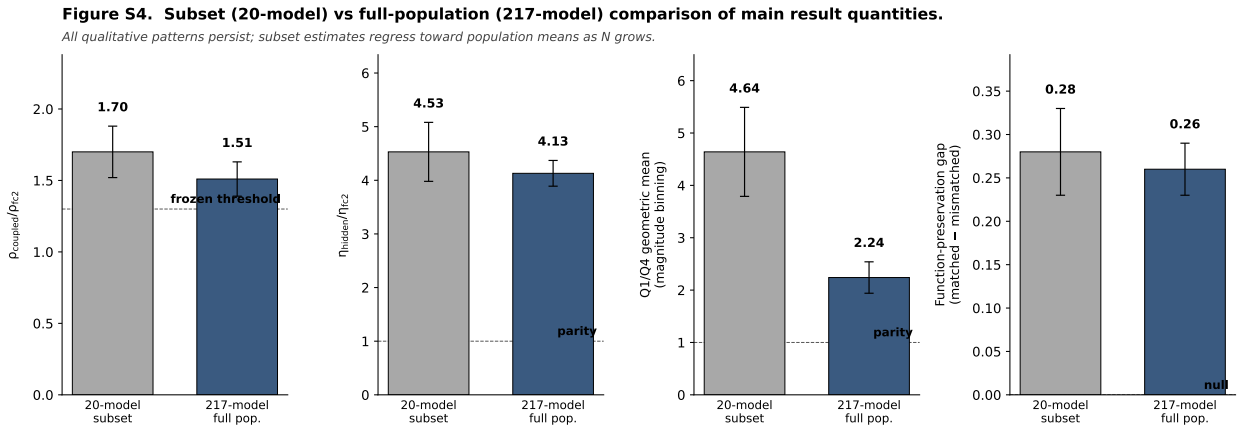

Figure 6: Subset (20-model) vs full-population (217-model) comparison of main result quantities. The 20-model subset is the paper-one continuity sample; the 217-model full population is the headline analysis. All qualitative patterns persist between the two sample sizes: coupling ratio remains above the frozen 1.30× threshold (1.70 → 1.51), η ratio remains above 4× (4.53 → 4.13), magnitude-binning Q1/Q4 geometric mean remains above 2× (4.64 → 2.24), and function-preservation gap remains around 0.27. Quantitative values regress toward population means as N grows — subset estimates are noisier point estimates from a smaller sample, not different effects in kind.

Table 1: Training grid for the 217-model grokked population. The population is constructed by sweeping the schedule grid below across 10 initialization seeds and retaining only runs that cleared the post-grokking test-accuracy threshold within the training budget.

Parameter	**Value**
Task	Modular addition over $\mathbb{Z}/47\mathbb{Z}$
Input space	$\{0, \ldots, 46\}^2$, $\lvert\text{input}\rvert = 2{,}209$
Train/test split	40%/60% ($\lvert\text{train}\rvert = 883$, $\lvert\text{test}\rvert = 1{,}326$), fixed across population
Architecture	2-layer MLP: embed (32) → hidden (128, ReLU) → logits (47)
Trainable parameters	15,887
Optimizer	AdamW, full-batch, cross-entropy loss
Training budget	24,000 steps
Grokking filter	test accuracy ≥ 0.95 within budget
Random seeds (10 total)	Original: $\{7, 13, 21, 42, 99\}$
	Extension: $\{101, 137, 211, 313, 401\}$
Constant schedules:	weight decay $\lambda \in \{0.3, 0.5, 1.0, 1.5, 2.0, 3.0\}$
	learning rate $\eta \in \{0.005, 0.01, 0.02\}$
Cyclic schedules:	$\lambda \in [0.3, 2.5]$ with periods $\{1500, 3000, 6000\}$ steps
	$\lambda \in [0.5, 2.0]$ with period 3000 steps
	learning rate $\eta = 0.01$ throughout
Attempted runs	232
Grokked runs (analyzed)	217
Failure mode of non-grokked runs	cluster in low-λ, low-η corner of grid

All training configurations and the per-seed/per-schedule grokking rates are reproduced from Dixon (2026, sec. 3). The 15 non-grokked runs are excluded from all population statistics, intervention experiments, and decoder fits reported in this paper.

Chan, Lawrence, Adrià Garriga-Alonso, Nicholas Goldowsky-Dill, Ryan Greenblatt, Jenny Nitishinskaya, Ansh Radhakrishnan, Buck Shlegeris, and Nate Thomas. 2022. "Causal Scrubbing: A Method for Rigorously Testing Interpretability Hypotheses." Alignment Forum / Redwood Research. https://www.alignmentforum.org/posts/JvZhhzycHu2Yd57RN/causal-scrubbing-a-method-for-rigorously-testing.

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*. Vol. 36. https://arxiv.org/abs/2304.14997.

Dixon, Kamil. 2026. "Grokking Collapses the Algorithm, Not the Function: Irrep-Energy Underdetermination in Modular Addition."

Geiger, Atticus, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, et al. 2023. "Causal Abstraction: A Theoretical Foundation for Mechanistic Interpretability." https://arxiv.org/abs/2301.04709.

Geiger, Atticus, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. "Causal Abstractions of Neural Networks." In *Advances in Neural Information Processing Systems*. Vol. 34. https://proceedings.neurips.cc/paper/2021/hash/4f5c422f4d49a5a807eda27434231040-Abstract.html.

Geiger, Atticus, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D. Goodman. 2024. "Finding Alignments Between Interpretable Causal Variables and Distributed Neural Representations." In *Proceedings of the Conference on Causal Learning and Reasoning*, 236:160–87. PMLR. https://arxiv.org/abs/2303.02536.

Goldowsky-Dill, Nicholas, Chris MacLeod, Lucas Sato, and Aryaman Arora. 2023. "Localizing Model Behavior with Path Patching." https://arxiv.org/abs/2304.05969.

He, Jianliang, Leda Wang, Siyu Chen, and Zhuoran Yang. 2026. "On the Mechanism and Dynamics of Modular Addition: Fourier Features, Lottery Ticket, and Grokking." https://arxiv.org/abs/2602.16849.

Heimersheim, Stefan, and Neel Nanda. 2024. "How to Use and Interpret Activation Patching." https://arxiv.org/abs/2404.15255.

Meng, Kevin, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. "Locating and Editing Factual Associations in GPT." In *Advances in Neural Information Processing Systems*. Vol. 35. https://arxiv.org/abs/2202.05262.

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*. https://arxiv.org/abs/2301.05217.

Power, Alethea, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. 2022. "Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets." https://arxiv.org/abs/2201.02177.

Truong, Xuan Khanh, Quynh Hoa Truong, Duc Trung Luu, and Thanh Duc Phan. 2026. "Spectral Entropy Collapse as an Empirical Signature of Delayed Generalisation in Grokking." https://arxiv.org/abs/2604.13123.

Turner, Alexander Matt, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. "Activation Addition: Steering Language Models Without Optimization." https://arxiv.org/abs/2308.10248.

Vig, Jesse, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. "Investigating Gender Bias in Language Models Using Causal Mediation

Analysis." In *Advances in Neural Information Processing Systems*. Vol. 33. https://proceedings.neurips.cc/paper/2020/hash/92650b2e92217715fe312e6fa7b90d82-Abstract.html.

Wang, Kevin, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. "Interpretability in the Wild: A Circuit for Indirect Object Identification in GPT-2 Small." In *International Conference on Learning Representations*. https://arxiv.org/abs/2211.00593.

Zhang, Fred, and Neel Nanda. 2023. "Towards Best Practices of Activation Patching in Language Models: Metrics and Methods." In *International Conference on Learning Representations*. https://arxiv.org/abs/2309.16042.

Zou, Andy, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, et al. 2023. "Representation Engineering: A Top-down Approach to AI Transparency." https://arxiv.org/abs/2310.01405.