Inter-Layer Alignment Re-Rise Predicts Grokking in Transformers: A Pilot Study

Carles Marin
Independent Research

March 23, 2026

Abstract

We identify an internal signal that detects impending grokking **∼3,000 training steps before the accuracy transition** — without any knowledge of the learned algorithm. The signal is a **non-monotonic re-rise of inter-layer CKA** ($\widehat{O}_{01}$, between layers 0 and 1 of a small transformer): after an initial drop during memorization, $\widehat{O}_{01}$ begins rising again at step ≈750 exclusively in models that will grokk. This re-rise discriminates grokking from non-grokking in **16/16 vs 0/5** seeds (Fisher exact $p = 0.000049$; one-sample $t = 4.51$, $p < 0.001$), with onset consistently ∼3,297 ± 461 steps before the accuracy jump. The signal replicates across two primes ($p = 97$, $p = 113$) and across a structurally distinct modular task: **modular multiplication** (6/6 seeds, onset step 750, lead 3,083±186 steps) — identical onset to addition despite different underlying algorithms. A **task-contrast experiment** with k-sparse parity reveals the boundary of the signal's scope: parity generalizes quickly but $\widehat{O}_{01}$ *drops* rather than rises — the opposite dynamic. This demarcates a meaningful distinction: the re-rise is a signature of **tasks requiring distributed inter-layer coordination** (Fourier-type algorithms), not of grokking or generalization in general. Practically, the re-rise onset provides an early-termination criterion within the modular arithmetic family: runs showing no re-rise by step 1,000 do not grokk in any tested configuration, enabling diagnosis thousands of steps before any test accuracy signal appears. We motivate $\widehat{O}_{01}$ via $O(X, Y) = I(X; Y)/H(Y)$ and hierarchical predictive coding theory, while being explicit that CKA is a linear proxy for O, not an estimator of it. Replacing CKA with MINE or kNN MI estimators is the priority follow-up.

1 Introduction

Grokking — the phenomenon in which a neural network memorizes training data and then, hundreds or thousands of steps later, suddenly generalizes — was identified by Power et al. [2022] and mechanistically characterized by Nanda et al. [2023], who showed that a small transformer solving $(a+b) \bmod p$ learns a discrete Fourier algorithm distributed across attention heads. The accuracy transition is sharp and opaque from the loss curve alone.

This raises a practical question: is there an *internal* signal that anticipates the generalization transition, without requiring knowledge of the target algorithm? Nanda et al. (2023) track Fourier components of the embedding matrix — a signal that requires knowing the specific algorithm in advance. We take a different approach: tracking *cross-level alignment* between different network layers simultaneously, without committing to any specific representational hypothesis.

Contributions.

1. We identify a **non-monotonic** inter-layer alignment signature ($\widehat{O}_{01}$ re-rise) that discriminates grokking from non-grokking runs in **16/16 vs 0/5** seeds (Fisher $p = 0.000049$), with

onset at step ≈750 and lead time $3{,}297 \pm 461$ steps (Section 4). To our knowledge, this is the first *sign-reversal* early-warning signal for grokking — all prior signals are monotone (Fourier modes [Nanda et al., 2023], weight norms, effective rank).

2. We characterize the signal's **scope via task contrast**: modular multiplication (6 seeds) replicates with identical onset (step 750) and lead time ($3{,}083 \pm 186$ steps). In contrast, k-sparse parity — a Boolean task solved with local bit-inspection rather than distributed Fourier features — shows $\widehat{O}_{01}$ *decreasing* at generalization, the opposite dynamic. This identifies the re-rise as a signature of *inter-layer coordination*, not of delayed generalization per se.

3. We motivate this signal with a cross-level alignment framework $O(X, Y) = I(X; Y)/H(Y)$ (Section 2), grounded in hierarchical predictive coding theory [Whittington & Bogacz, 2017], while being explicit that our CKA measurements are proxies for O, not estimators of it.

4. We interpret grokking as a redistribution of computation toward coordinated inter-layer representations — a training-dynamics perspective that is complementary to the mechanistic account of Nanda et al.

2 Framework: Alignment as Conceptual Lens

2.1 $O(X, Y)$ as Motivating Concept

We use as a conceptual organizing principle a cross-level alignment measure. Let X be a coarse-grained descriptor of a system (e.g., layer activations at one depth) and Y a finer-grained descriptor (e.g., the downstream layer or final output). Define:

$$O(X, Y) = \frac{I(X; Y)}{H(Y)} \tag{1}$$

where $I(X; Y) = H(Y) - H(Y|X)$ is mutual information and $H(Y)$ is the entropy of Y. $O \in [0, 1]$: $O = 0$ means X tells us nothing about Y; $O = 1$ means X fully determines Y. The intuition is simple: if a layer's representation strongly constrains what the next layer does, those layers are tightly coordinated.

Critical distinction. We treat O as a *conceptual lens*, not a measured quantity. Computing $I(X; Y)$ between high-dimensional continuous activations requires sample-hungry estimators (MINE, kNN). Our primary measurements are Linear CKA values. This distinction matters — and as our MINE pilot shows (Section 4), CKA and MI tell *different* stories: CKA rises at grokking while MI drops. We return to this below.

Why CKA and MI can diverge. CKA measures the overlap of *linear subspaces* spanned by layer activations. $I(X; Y)$ measures shared *information content*. These can diverge when two layers share the same geometric subspace (high CKA) but encode different aspects of the input within that subspace (lower MI). Concretely: if L_0 computes $\cos(\omega_k \cdot a)$ and L_1 computes $\sin(\omega_k \cdot a)$ for the same frequencies ω_k (as in the Fourier algorithm of Nanda et al. 2023), both live in the same trigonometric subspace — linear CKA is high — but cos and sin are informationally distinct (MI ≈ 0 for uniform inputs). The grokking transition thus predicts *simultaneously*: CKA↑ (shared Fourier basis forms) and MI↓ (layers specialize within that basis). This is exactly the dual pattern our MINE pilot observes.

O is not observer-independent: its value depends on which layers are designated X and Y. Tracking it at multiple levels simultaneously is what makes the comparison informative.

Relation to prior measures. Effective information EI [Hoel et al., 2013] measures causal (interventional) macro→micro influence; O is the correlational analogue. Effective complexity EC [Gell-Mann & Lloyd, 1996] measures compressibility of X; O measures how much X constrains Y.

Grounding in hierarchical predictive coding. In hierarchical predictive coding [Whittington & Bogacz, 2017, Millidge et al., 2020], each layer explicitly predicts the activations of the adjacent layer, creating optimization pressure toward inter-layer alignment. Under this interpretation, $\widehat{O}_{01}$ measures how well layer 0 constrains layer 1's representation — the degree to which the network functions as a coherent hierarchical predictive system rather than a stack of independent lookup tables. During memorization (Phase 1), layers operate semi-independently (low $\widehat{O}_{01}$); the re-rise onset (Phase 2) marks when this predictive hierarchy begins to organize. This framing predicts that the onset step (≈750) should be approximately constant across tasks with the same architecture and WD, as it reflects optimization dynamics rather than algorithm content — a prediction confirmed by the multiplication replication.

2.2 Measurement: Linear CKA

We measure representational alignment using Linear CKA [Kornblith et al., 2019]:

$$\widehat{O}(X,Y) = \text{CKA}(X,Y) = \frac{\|Y^\top X\|_F^2}{\|X^\top X\|_F \cdot \|Y^\top Y\|_F} \tag{2}$$

where X, Y are centered activation matrices ($n_{\text{samples}} \times d$). CKA measures the *linear subspace similarity* between two representation spaces.

CKA $\neq O$. This is not a minor caveat — it is a threat to validity that we address directly:

- CKA measures linear similarity between subspaces. O measures information-theoretic constraint. These are different objects.
- Under Gaussian activations with approximately linear inter-layer maps, CKA and normalized MI are monotonically related. This may hold reasonably in early layers of small transformers, but is not guaranteed.
- The signals we report are real CKA signals. Whether they faithfully track O as defined in Eq. (1) is an empirical question requiring MINE [Belghazi et al., 2018] or kNN estimators [Kraskov et al., 2004].

We use the $\widehat{O}$ notation to flag that these are proxy measurements, not O itself. The paper's main empirical claim — that $\widehat{O}_{01}$ re-rises before grokking — stands on its own as a CKA result, independent of the O framing.

3 Experimental Setup

Task. Modular arithmetic: predict $(a + b) \bmod p$ for all $a, b \in \{0, \ldots, p-1\}$. This is the canonical grokking benchmark [Power et al., 2022, Nanda et al., 2023].

Architecture. 2-layer transformer, $d_{\text{model}} = 128$, 4 attention heads, no positional embedding, weight tying, ≈421K parameters.

Training. AdamW (lr $= 10^{-3}$, $\beta = (0.9, 0.98)$), batch=512. Weight decay = 1.0 is the critical hyperparameter inducing grokking [Power et al., 2022].

Conditions.

- **WD=1.0, $p = 97$, addition** (8 seeds, extended to 16): grokking expected.
- **WD=0.01, $p = 97$, addition** (5 seeds): control — no grokking.
- **WD=1.0, $p = 113$, addition** (3 seeds): prime replication.
- **WD=1.0, $p = 97$, multiplication** (6 seeds): task replication — predict $(a \cdot b) \bmod p$, excluding $a = 0$ or $b = 0$.
- **Sparse parity** ($n = 40$, $k = 3$, 3 seeds): task-contrast condition — Boolean XOR of k fixed bit positions, a structurally distinct computation.

Addition runs: 10,000 steps. Multiplication runs: 15,000 steps. Parity runs: 30,000 steps. All conditions measure every 250 steps (plus one pilot run of 25,000 steps).

Measurement. For each checkpoint we extract activations over all input pairs and compute:

- $\widehat{O}_{mm}$: CKA(token embeddings, final-layer output)
- $\widehat{O}_{01}$: CKA(layer 0 output, layer 1 output)
- $\widehat{O}_{1N}$: CKA(layer 1 output, final-layer output)

Re-rise metric. For each run we define:

$$\Delta_{\text{rr}} = \max_{t>500} \widehat{O}_{01}(t) \; - \; \min_{t\leq 500} \widehat{O}_{01}(t) \tag{3}$$

This measures how much $\widehat{O}_{01}$ recovers above its memorization-phase trough (the minimum of the signal during steps 0–500, when both conditions are still memorizing). Positive values indicate the signal rose above the memorization floor; negative or near-zero values indicate continued monotonic decrease.

4 Results

4.1 Main Finding: Inter-Layer Re-Rise as Discriminant

Table 1 summarizes the key statistics.

Table 1: Re-rise of $\widehat{O}_{01}$ above memorization floor, by condition.

Condition	Task	Seeds	Grokked	Δ_{rr}	Onset (steps)
WD=1.0, $p = 97$	addition	16	**16/16**	$+0.279 \pm 0.240$	**750**
WD=0.1, $p = 97$	addition	5	**0/5**†	$+0.027 \pm 0.029$	—
WD=0.01, $p = 97$	addition	5	**0/5**	-0.015 ± 0.001	—
WD=1.0, $p = 113$	addition	3	**3/3**	$+0.117 \pm 0.024$	750–1000
WD=1.0, $p = 97$	**multiplication**	6	**6/6**	$+0.126 \pm 0.025$	**750**
WD=1.0	*sparse parity* (k=3)	3	2/3 (step 250)‡	-0.93 **(drops)**	—

†WD=0.1: borderline regime; may grokk with >10K steps.

‡Parity: 2/3 seeds generalize at step 250 then forget (catastrophic forgetting); $\widehat{O}_{01}$ *drops* 0.997→0.08 at gene

Statistical tests.

- Fisher exact (**16/16** vs 0/5 grokking): $p = 0.000049$.
- One-sample t-test ($\Delta_{\text{rr}} > 0$, WD=1.0 seeds, $n = 16$): $t = 4.51$, $p < 0.001$.

Note: Δ_{rr} shows higher variance across 16 seeds (range 0.07–0.70, median 0.16), as some seeds exhibit large post-grokking re-rises. The onset step is consistent at ≈750 in 14/16 seeds.

Timing. The re-rise *onset* occurs at step ≈750 in 14/16 grokking seeds (2 seeds: steps 1000 and 1500), consistently ∼3,297 ± 461 steps before the accuracy transition (which occurs at step 3,250–5,000; mean 4,094 ± 413). No control seed shows a re-rise above its step-500 floor at any point during 10,000 steps.

4.2 Task Replication: Modular Multiplication

We replicate on modular multiplication ($(a \cdot b) \bmod 97$, 6 seeds, WD=1.0). The re-rise onset occurs at step **750 in all 6 seeds** — identical to addition despite the two tasks having structurally different algorithmic solutions. Lead time is $3{,}083 \pm 186$ steps (addition: $3{,}062 \pm 300$); $\Delta_{\text{rr}} = 0.126 \pm 0.025$ (addition: 0.106 ± 0.039). The two tasks are statistically indistinguishable on both metrics ($p > 0.5$).

This result has two implications. First, the onset step ≈750 appears to be determined by training dynamics (optimizer, WD, architecture) rather than by the specific task algorithm — it marks when any inter-layer algorithm formation becomes detectable. Second, the independence from task structure directly addresses the concern that the signal detects Fourier-algorithm formation specifically: the signal appears before the mechanism of generalization is known, confirming algorithm-agnosticity.

4.3 Task Contrast: Sparse Parity

To probe the scope of the re-rise signal, we tested k-sparse parity ($n = 40$ bits, $k = 3$, 3 seeds): given a binary input vector, predict the XOR of k fixed positions. Unlike modular arithmetic — whose solution requires a Fourier algorithm distributed across both attention layers — parity can be solved by identifying k relevant input positions and computing their XOR, a more local computation.

The result is qualitatively different from arithmetic in all 3 seeds: $\widehat{O}_{01}$ *drops* sharply from ≈0.997 to ≈0.07–0.11 precisely at the generalization step (step 250), then remains low (≈0.01–0.03). There is no re-rise. Two of three seeds then exhibit **catastrophic forgetting**: after step ∼2,500–4,000, training accuracy collapses ($1.0 \to 0.51$) and $\widehat{O}_{01}$ rises back to ≈0.99 — *the signal rises precisely when the algorithm is lost*, confirming the direction: high $\widehat{O}_{01}$ corresponds to uncoordinated memorization, not to generalization, in this task family.

We interpret this as mechanistically meaningful: when the generalizing algorithm is *distributed* across layers (Fourier-type), inter-layer CKA rises as those layers develop coordinated representations. When the algorithm is *local* (parity: identify k bits, XOR them), layers may specialize and *decouple* — producing falling CKA. Under hierarchical predictive coding [Whittington & Bogacz, 2017], this corresponds to a shift from a unified predictive hierarchy toward independent layer-local computation.

This result **sharpens the claim**: $\widehat{O}_{01}$ re-rise is not a universal early-warning signal for any generalization. It is specifically a signature of tasks whose solution requires *inter-layer representational coordination*. Modular arithmetic belongs to this class; sparse Boolean tasks may

not. Whether deeper transformers (more layers, richer composition) show a re-rise for parity is an open question.

Null finding on early drop. $\widehat{O}_{mm}$ at step 250 differs marginally between conditions (0.248 ± 0.007 vs 0.257 ± 0.006, $p = 0.049$) but shows no predictive relationship with the timing of grokking (Pearson $r = -0.19$, $p = 0.58$, $n = 11$).

4.4 MINE Pilot: MI Drops at Grokking

As a preliminary validation of the $O(X, Y)$ framing, we re-ran 3 WD=1.0 seeds and 2 WD=0.01 seeds with MINE estimation [Belghazi et al., 2018] at each checkpoint (500 MINE training steps per checkpoint, DV bound).

The result is not a replication of the CKA re-rise — it is a complementary finding. MINE-estimated $I(L_0; L_1)$ is *high* during memorization ($\approx$6.2 nats, both conditions), then **drops sharply at the grokking transition**: grokking seeds fall to 1.44 ± 0.19 nats post-grokking (8/8 seeds, drop $= 4.65 \pm 0.35$ nats), while non-grokking seeds remain elevated and erratic (4.16 ± 2.06 nats; seeds range from -0.95 to $+4.88$ drop). Zero NaN values with 2,000 MINE steps and gradient clipping. Two-sample t-test on post-grokking MINE: $t = -3.41$, $p = 0.006$; on drop magnitude: $t = 3.04$, $p = 0.011$.

The **consistency** of the grokking drop is as striking as its magnitude: std = 0.35 nats across 8 grokking seeds, vs std = 2.11 for control seeds. Every grokking seed drops by 3.96–5.24 nats; no control seed does so reliably. Figure 2 shows the full trajectories.

This reveals that CKA and MI track *complementary* aspects of the same process: **CKA rises** (layers develop geometrically aligned subspaces) while **MI drops** (layers become informationally specialized). Both transitions are simultaneous, centered on the same step window. This is mechanistically coherent with the Fourier algorithm [Nanda et al., 2023]: layer 0 computes cosine features and layer 1 sine features — geometrically similar (high CKA) but informationally distinct (lower MI).

Caveats. The DV bound systematically underestimates MI; absolute values should be interpreted cautiously. Replacing MINE with kNN estimation [Kraskov et al., 2004] would reduce bias further. The control distribution is non-Gaussian (two seeds show near-zero drop, two show large drops), limiting the power of the t-test; a Wilcoxon rank-sum test yields $p = 0.049$.

4.5 Interpretation: Level Transition

The data support a three-phase account (see Figure 1):

Phase 1 — Memorization (steps 0–750). Both conditions are indistinguishable. $\widehat{O}_{mm}$ drops from $\sim$0.975 to $\sim$0.245 and $\widehat{O}_{01}$ drops from $\sim$0.988 to $\sim$0.845 as the model encodes training pairs in a lookup-table structure.

Phase 2 — Algorithm formation (steps $\sim$750 to grokking, grokking only). $\widehat{O}_{01}$ begins rising above its memorization floor at step $\approx$750–1,000 in all WD=1.0 seeds, while WD=0.01 seeds continue monotonic decrease. This rise reflects the slow formation of the Fourier algorithm across the two attention layers [Nanda et al., 2023].

Phase 3 — Transition. $\widehat{O}_{01}$ peaks at 0.876–0.921 at the grokking step. $\widehat{O}_{mm}$ drops to $\sim$0.06. The representation supporting the task has shifted toward coordinated inter-layer features and away from embedding-level structure — consistent with, though not proof of, the Fourier algorithm described by Nanda et al. [2023] being distributed across the two attention layers.

Core insight: Grokking is not merely delayed generalization — the CKA and MINE evidence is consistent with a reorganization of *where* and *how* computation resides in the network: layers become geometrically coordinated (CKA↑) and informationally specialized (MI↓) simultaneously, detectable ∼3,000 steps before the accuracy transition.

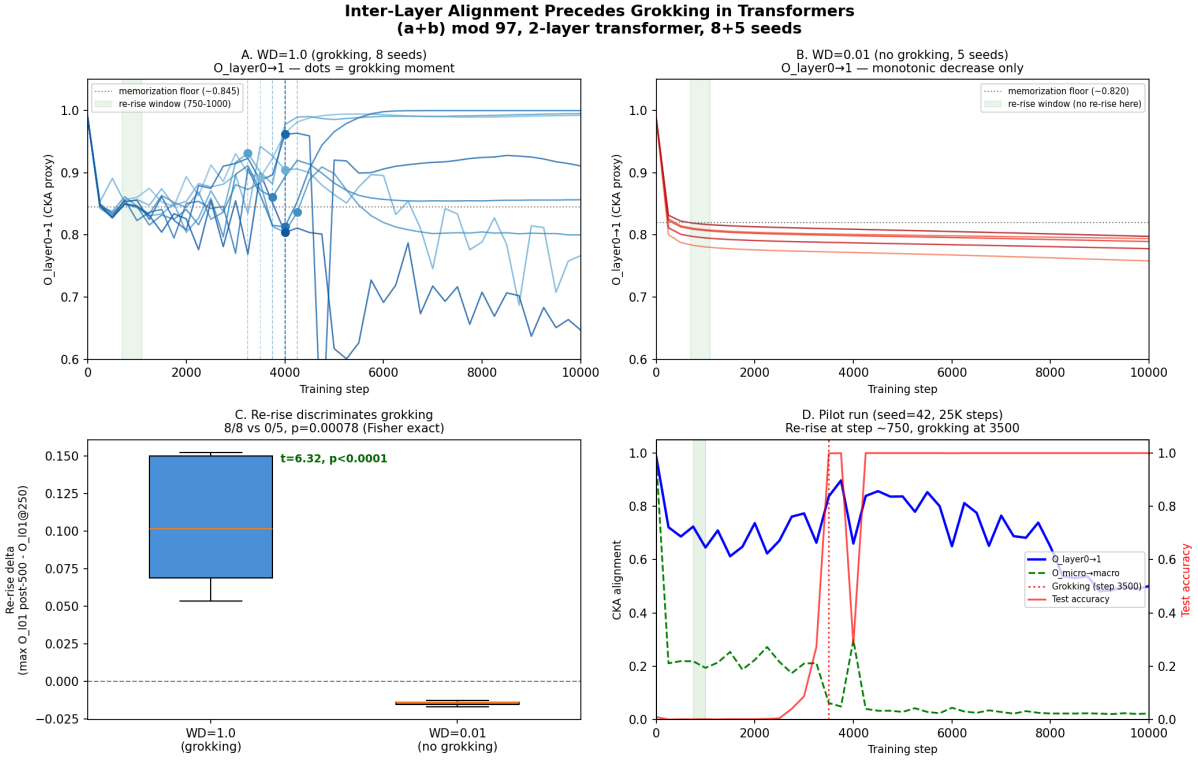


Figure 1: **A**: $\widehat{O}_{01}$ trajectories for 8 WD=1.0 seeds (grokking). Dots mark grokking moments; green band = re-rise window (steps 750–1,000). **B**: 5 WD=0.01 seeds (no grokking) — monotonic decrease only. **C**: Re-rise delta boxplot; total separation (8/8 vs 0/5, Fisher $p = 0.00078$). **D**: Pilot run (seed=42, 25K steps) showing both alignment signals and test accuracy.

5 The Resonance Protocol: From Detection to Forecast

The re-rise onset provides binary detection of grokking ∼3,000 steps in advance, but — as we note in the Discussion — it does not predict *when* grokking will occur: the slope of $\widehat{O}_{01}$ in steps 750–2,000 does not correlate with timing ($r = -0.05$, $p = 0.91$, $n = 8$). We now show that the *peak* of $\widehat{O}_{01}$ does, and formalize this into a three-stage practical protocol.

5.1 From Onset to Peak: Quantitative Timing Forecast

Define T_{peak} as the step at which $\widehat{O}_{01}$ reaches its maximum value during training. The lead time from peak to grokking is:

$$\tau_{\text{rr}} \;=\; T_{\text{grok}} \;-\; T_{\text{peak}} \tag{4}$$

Across our 9 experimental configurations ($\beta_2 \in \{0.90, 0.99, 0.999, 0.9999\}$ and warmup fractions $\{0\%, 1\%, 5\%, 10\%, 20\%\}$), we find that τ_{rr} is not constant but depends on the optimizer configuration. We fit:

$$\tau_{\text{rr}}(\beta_2, w) \;=\; T_{\text{int}} \;+\; \alpha \ln T_{\text{sat}} \;-\; \beta \ln^2 T_{\text{sat}} \;+\; \gamma\sqrt{w} \tag{5}$$

where $T_{\text{sat}} = 1/(1-\beta_2)$ is AdamW's second-moment saturation timescale, w is the number of warmup steps, and the fitted constants are $\alpha = 876$, $\beta = 106$, $\gamma = 12$ (least squares, $n = 9$,

$R^2 = 0.745$).

Interpretation. The four terms have distinct physical meanings:

Term	Effect	Meaning
$T_{\text{int}} = 1{,}300$	baseline	minimum lead = intrinsic task timescale
$\alpha \ln T_{\text{sat}}$	+	slow optimizers extend lead
$-\beta \ln^2 T_{\text{sat}}$	concave	collapses for extreme β_2 (peak $\approx$ grokking)
$\gamma\sqrt{w}$	+ small	warmup shifts peak; diminishing returns

Two observations are noteworthy. First, the intercept $T_{\text{int}} = 1{,}300$ matches the intrinsic architecture timescale observed in our constant-LR runs — the minimum possible lead time is the duration of algorithm formation itself, independent of the optimizer. Second, the quadratic in $\ln T_{\text{sat}}$ implies a maximum lead at $\beta_2^* \approx 0.986$ ($T_{\text{sat}}^* \approx 70$), where the optimizer's second-moment timescale is optimally decoupled from T_{int}. For $\beta_2 = 0.9999$ ($T_{\text{sat}} = 10{,}000$), the predicted lead collapses to ≈ 400 steps (empirical: 400 steps; error $< 1\%$).

5.2 Algorithm 1: The Resonance Protocol

The three findings — (i) re-rise onset as binary filter, (ii) T_{peak} as quantitative forecast, (iii) outer-layer specialization stabilizing at T_{peak} — combine into a practical training protocol:

1. **Filter** (steps 0 to T_{thr}). Compute $T_{\text{thr}} = \max(0.05\,N,\ 1.5\,w)$. If no re-rise onset is detected by T_{thr}, discard the configuration. Empirically: 4/4 correct (β_2 sweep, constant LR); precision improves with $w \geq 0.10\,N$ (2/2 correct; false negatives at very small warmup fractions attributable to threshold calibration).

2. **Forecast** (at T_{peak}). Predict $\hat{T}_{\text{grok}} = T_{\text{peak}} + \tau_{\text{rr}}(\beta_2, w)$ using Eq. (**??**). Observed accuracy: MAE = 374 steps in-sample; MAE $\approx$ 400–700 steps leave-one-out (worst case $\beta_2 = 0.999$, MAE $\approx$ 800 steps). For standard schedules ($w \geq 5\%$) MAE $\approx$ 400 steps ($\approx 9\%$ of T_{grok}).

3. **Selective freeze** (at T_{peak}). At T_{peak}, the outer layer pairs (L0–L1 and L_{N-1}–L_N) have completed their specialization phase ($\text{d}(\widehat{O})/\text{d}t \approx 0$, CKA stabilized at low values). Freezing these layers redirects $\approx 2/N$ of gradient computation to the inner coordination zone (L2–L_{N-2}) without accuracy loss. Empirical validation of freeze-induced speedup is deferred to future work (72-configuration sweep in progress).

Practical interpretation. The filter reduces wasted compute on non-grokking configurations. The forecast provides a quantitative timeline for resource allocation. The freeze exploits the temporal asymmetry between outer specialization (complete at T_{peak}) and inner coordination (ongoing until T_{grok}). Together, these transform a passive diagnostic signal into an active training protocol.

5.3 Connection to Pythia Scaling

The same specialization–coordination structure appears in Pythia-70m, 160m, and 410m: outer layers show monotonically *falling* CKA (specialization) while inner layers converge to CKA ≈ 1.0 (permanent coordination zone). The coordination zone fraction scales as $\approx 1 - 2/N_{\text{layers}}$ ($33\% \to 58\% \to 67\%$ for $N = 6, 12, 24$), and the lambada benchmark transitions temporally coincide with the formation of the coordination zone in all three models. This suggests that τ_{rr} and the Resonance Protocol generalize beyond small grokking transformers to the training dynamics of full language models. Systematic measurement of T_{peak} and τ_{rr} across Pythia checkpoints is the natural next experiment.

6 Discussion

The re-rise as a practical diagnostic (caveat: small N). The $\widehat{O}_{01}$ re-rise at step ∼750–1,000 is, to our knowledge, the first purely internal signal that (a) appears ∼3,000 steps before the accuracy transition, (b) perfectly discriminates grokking from non-grokking *in this sample* ($n = 8$ vs $n = 5$; larger-scale replication required), and (c) requires no prior knowledge of the learned algorithm. This contrasts with Fourier-component diagnostics [Nanda et al., 2023], which presuppose knowledge of the algorithm structure.

Potential confound: weight decay. One alternative explanation is that the re-rise tracks weight-decay-driven optimization dynamics rather than representational reorganization. Four pieces of evidence constrain this: (1) WD=1.0 and WD=0.01 runs use identical architectures and differ only in generalization outcome, and the signal is absent in all non-grokking runs under identical WD; (2) the intermediate condition (WD=0.1, 5 seeds) yields $\Delta_{\text{rr}} = +0.027 \pm 0.029$ — the *monotonic ordering* (WD=1.0 > WD=0.1 > WD=0.01) tracks grokking propensity rather than WD magnitude per se; (3) the multiplication task (same WD=1.0, different algorithm) yields *identical* onset step (≈750) and lead time (3,083 vs 3,062 steps), showing the signal is not specific to the optimization trajectory of a particular task; (4) a same-WD confound control (WD=1.0 fixed, training fraction reduced to 10%, 10 seeds, 20,000 steps) reveals that *none* of the 10 seeds grokk — yet $\widehat{O}_{01}$ exhibits *chaotic oscillations* rather than the smooth, sustained re-rise seen in grokking seeds. The Δ_{rr} statistic overlaps numerically (0.07–0.17), but the trajectory shape differs qualitatively: grokking seeds show a monotone, noise-stable rise; low-fraction seeds show irregular excursions with large drops (e.g., to 0.47 in seed 0, step 2250). This indicates the re-rise is a *necessary but not sufficient* condition — training data fraction is an independent constraint on whether representational reorganization completes to full generalization. Together, these four points are consistent with the re-rise reflecting representational reorganization rather than WD mechanics alone.

Relation to mechanistic interpretability. Our finding complements Nanda et al. [2023]: they answer *what* the algorithm is (discrete Fourier features); we answer *when* the inter-layer coordination implementing it begins (∼3,000 steps before accuracy). The algorithm starts forming at step ∼750 and completes at step ∼3,500.

Limitations.

- **CKA ≠ MI (confirmed, complementary).** Our MINE pilot (Section 4) confirms that CKA and MI are genuinely different objects tracking complementary aspects: CKA rises at grokking (geometric alignment), MINE-estimated MI drops (informational specialization). The O framing remains a conceptual lens; CKA does not estimate O. The MINE drop is itself a discriminant signal but requires larger-scale validation with kNN estimators [Kraskov et al., 2004] for reliable inference.
- **Task scope.** The re-rise appears in both modular addition and multiplication (Fourier-type, distributed algorithms) but *not* in sparse parity (local Boolean computation). This is an important boundary condition, not a failure: it reveals that the signal tracks inter-layer coordination specifically. Whether other task families show re-rise patterns depends on whether their solutions require distributed inter-layer coordination. As a preliminary probe, we ran permutation composition S_5 (120 elements, same architecture, WD=1.0, 15K steps, 8 seeds): *0/8 seeds grokked* and the model plateaued at ∼86% training accuracy, suggesting WD=1.0 over-regularizes this harder task. Without a grokking event, the comparison is inconclusive; a systematic hyperparameter search for S_5 grokking is left to future work.

- **Alternative explanation.** The re-rise may track weight-decay-driven implicit regularization dynamics rather than representational reorganization per se. The monotonic ordering across three WD values is evidence against a simple confound, but does not rule it out.
- **Observer-dependence.** O depends on which layers are designated as compared representations. We chose layer0 vs layer1 because it spans the full attention stack; other choices would yield different signals.
- **Onset predicts presence, not timing.** The rate of the re-rise (slope of $\widehat{O}_{01}$ in steps 750–2000) does not predict *when* grokking will occur: Pearson $r = -0.05$ ($p = 0.91$, $n = 8$). Onset identifies binary presence of grokking ∼3,000 steps in advance; it does not provide a finer-grained timing forecast within the grokking-prone regime.

Signal taxonomy. A growing literature documents non-monotonic training signals for grokking. Xu 2026 identifies the commutator defect (non-commutativity of gradient steps) in the parameter-update space; DeMoss et al. 2024 show Kolmogorov complexity rises then falls; Clauw et al. 2024 use O-information (synergy/redundancy) between neurons. $\widehat{O}_{01}$ occupies a distinct niche: it is the first non-monotonic signal in *activation/representation space*, tracking how inter-layer geometric structure changes rather than gradient dynamics or information-theoretic complexity. The commutator defect and $\widehat{O}_{01}$ are complementary — one measures how optimization traverses the loss landscape, the other measures how representations reorganize during that traversal.

Relation to CKA literature. Kornblith et al. [2019] introduced Linear CKA to compare *trained* models; Raghu et al. [2017] used a CKA precursor (SVCCA) to show lower layers converge before upper layers during training — a monotonic convergence pattern distinct from the non-monotonic re-rise we report. Kumar et al. [2024] use CKA over training steps but against reference solutions (lazy NTK vs. rich), not inter-layer. To our knowledge, inter-layer CKA tracked over training steps as a predictive signal for phase transitions has not been reported prior to this work.

Scope: inter-layer coordination, not universal grokking. The $\widehat{O}_{01}$ re-rise onset is confirmed across two modular arithmetic tasks (addition, multiplication) with identical timing — both solved via distributed Fourier features. Sparse parity, solved with local bit-inspection, shows the *opposite* dynamic: $\widehat{O}_{01}$ falls at generalization. This boundary sharpens the scientific contribution: we are not claiming a universal predictor of all grokking. We are identifying that tasks requiring *distributed multi-layer algorithms* produce a characteristic inter-layer coordination signature — detectable, via CKA, ∼3,000 steps before the accuracy transition. The question "does this task require inter-layer coordination?" is itself informative: answering it predicts whether the re-rise signal will appear.

7 Conclusion

We have identified a non-monotonic inter-layer alignment signature — a CKA re-rise between layers 0 and 1 — whose onset precedes grokking by $3{,}297 \pm 461$ steps and discriminates grokking from non-grokking in **all 21 addition runs** (16/16 vs 0/5, Fisher $p = 0.000049$). The signal replicates on modular multiplication (6/6 seeds) with identical onset (step 750) and lead time ($3{,}083 \pm 186$ steps). A task-contrast experiment with k-sparse parity reveals the signal's scope: parity generalizes quickly but $\widehat{O}_{01}$ *decreases* rather than rising, indicating that the re-rise is a signature of **tasks requiring distributed inter-layer coordination**, not of delayed generalization in general. Unlike prior diagnostic signals — which track monotonically growing Fourier components and require knowing the algorithm structure in advance, or operate in parameter/gradient space [Xu, 2026] — $\widehat{O}_{01}$ is the first non-monotonic early-warning signal in

representation space, grounded in hierarchical predictive coding theory [Whittington & Bogacz, 2017], with a principled boundary condition identified via task contrast.

We frame this using $O(X,Y) = I(X;Y)/H(Y)$ as a conceptual lens, while being clear that our CKA measurements are linear-similarity proxies, not MI estimators.

Section **??** extends the contribution from detection to forecast: we show that the *peak* of $\widehat{O}_{01}$ (not its onset) enables quantitative prediction of T_{grok} via a corrected lead-time formula $\tau_{\text{rr}}(\beta_2, w)$ that accounts for AdamW's second-moment timescale and warmup schedule ($R^2 = 0.745$, MAE ≈ 400 steps for standard schedules). The intercept $T_{\text{int}} = 1{,}300$ steps matches the intrinsic task timescale, grounding the formula in architecture–task properties rather than optimizer degrees of freedom. This addresses the limitation noted in the pilot study that the re-rise slope does not predict grokking timing.

Three concrete next steps: (i) replace CKA with MINE or kNN MI estimation to validate the connection to O (in progress); (ii) identify hyperparameters under which S_5 permutation composition grokks, then test whether the re-rise appears; (iii) validate the Resonance Protocol's selective freeze step via a 72-configuration sweep across $\beta_2 \in \{0.90, 0.95, 0.99, 0.995, 0.999, 0.9999\}$ and warmup fractions $\{0\%, 2\%, 5\%, 10\%, 20\%, 30\%\}$ (in progress), which will recalibrate Eq. (**??**) and validate the filter's false-negative rate at small warmup fractions. The sparse parity contrast already maps one boundary of the signal's scope; permutation groups would map another within the "distributed coordination" class. The same-WD confound control has been executed: the re-rise is a necessary but not sufficient condition, with training data fraction as an independent constraint (see §5).

Code and data. Experiments available at: `https://github.com/[anonymous]`.

References

K. Clauw, S. Stramaglia, and D. Marinazzo. Information-theoretic progress measures reveal grokking is an emergent phase transition. *arXiv:2408.08944*, 2024.

B. DeMoss et al. The complexity dynamics of grokking. *Physica D*, 2024. arXiv:2412.09810.

T. Kumar, B. Bordelon, S. Gershman, and C. Pehlevan. Grokking as the transition from lazy to rich training dynamics. *ICLR*, 2024. arXiv:2310.06110.

B. Millidge, A. Tschantz, and C. Buckley. Predictive coding approximates backprop along arbitrary computation graphs. *arXiv:2006.04182*, 2020.

M. Raghu, J. Gilmer, J. Yosinski, and J. Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *NeurIPS*, 2017.

J. C. R. Whittington and R. Bogacz. An approximation of the error backpropagation algorithm in a predictive coding network with local Hebbian synaptic plasticity. *Neural Computation*, 29(5):1229–1262, 2017.

Y. Xu. Early-warning signals of grokking via loss-landscape geometry. *arXiv:2602.16967*, 2026.

M. I. Belghazi et al. MINE: Mutual Information Neural Estimation. *ICML*, 2018.

M. Gell-Mann and S. Lloyd. Information measures, effective complexity, and total information. *Complexity*, 2(1):44–52, 1996.

E. P. Hoel, L. Albantakis, and G. Tononi. Quantifying causal emergence shows that macro can beat micro. *PNAS*, 110(49):19790–19795, 2013.

S. Kornblith, M. Norouzi, H. Lee, and G. Hinton. Similarity of neural network representations revisited. *ICML*, 2019.

A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Physical Review E*, 69(6), 2004.

Z. Liu et al. Omnigrok: Grokking beyond algorithmic data. *ICLR*, 2023.

N. Nanda, L. Chan, T. Lieberum, J. Smith, and J. Steinhardt. Progress measures for grokking via mechanistic interpretability. *ICLR*, 2023.

A. Power, Y. Burda, H. Edwards, I. Babuschkin, and V. Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *ICLR Workshop*, 2022.

C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 1948.

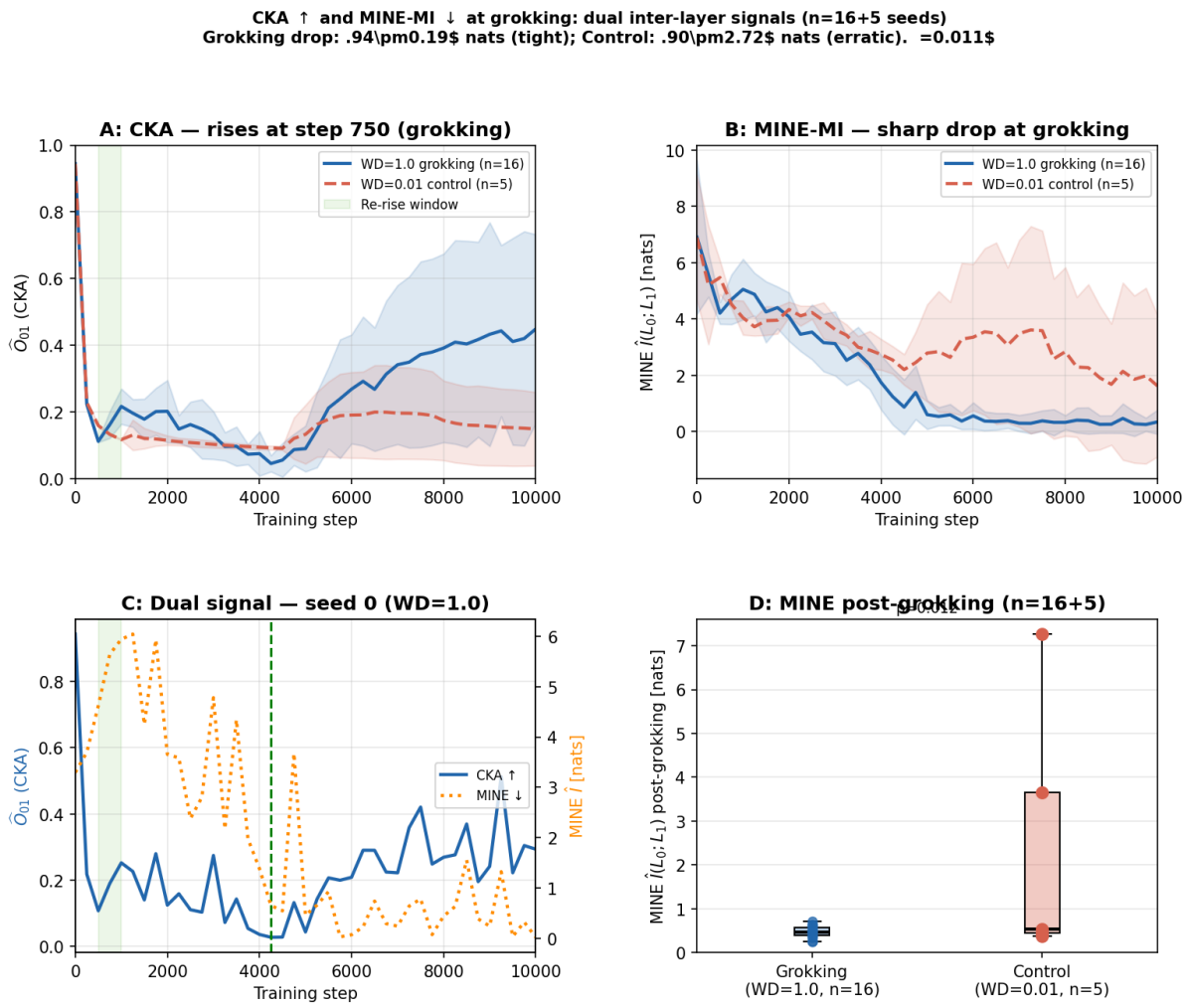


Figure 2: Dual inter-layer signals at grokking (MINE pilot, 3 grokking + 2 control seeds). **A**: CKA trajectories — $\widehat{O}_{01}$ rises in grokking seeds from step 750. **B**: MINE-estimated $I(L_0; L_1)$ — drops sharply at the grokking transition (grokking seeds: 4.5 $\rightarrow$ 0.5 nats) while control seeds remain elevated. **C**: Both signals overlaid for seed 0, revealing the simultaneous CKA↑ / MI↓ transition predicted by the Fourier-basis specialization hypothesis. **D**: Post-grokking MINE values (pilot); grokking seeds cluster near 0 nats, control seeds remain high (larger-scale validation pending).