

Subspace Crystallization in Grokking: Rank Compression Tracks the Phase Transition

Shlok Mehendale*
Lossfunk

Paras Chopra
Lossfunk

Abstract

Grokking describes understanding something deeply and intuitively. In neural networks, it is the phenomenon where a network generalizes long after memorizing its training data. This delay has been linked to a fall in the effective rank of the network's weight matrices, but whether this rank compression drives grokking or follows it is still unresolved. We study a mean-field MLP and a one-layer transformer trained on modular arithmetic and observe that: (i) *Compression follows the onset, not memorization.* Every transformer memorizes its training data within 250 steps, while the grokking onset (the first step at which test accuracy reaches 95%) varies from 6,500 to 21,000 steps. The steepest fall in effective rank comes about 900 steps after the onset, however late the onset occurs, and networks that memorize without grokking compress far less. (ii) *Subspace crystallization.* We call the span of the fewest top singular directions of a weight matrix that preserve 99% of test accuracy its *sufficient subspace*. Its size can shrink as the margin between the correct and the best wrong output grows, so we also count the directions needed to keep the outputs within 10%. In the transformer, this count falls during the transition from a quarter or more of the available directions to about a tenth, which we call *subspace crystallization*. (iii) *Imposing low rank does not speed up grokking.* Against a factorized full-rank control, capping the rank of the transformer's MLP input weights delays the onset, and capping the rank of its embedding stops 15 of 20 runs from grokking. Compression is therefore a *signature* of grokking rather than its cause.

1 Introduction

Grokking [Power et al., 2022], memorization followed by delayed generalization, offers a rare window into how circuits emerge from gradient descent: a clean phase transition from a lazy, memorizing regime to a rich, generalizing one [Kumar et al., 2024]. Substantial progress exists on *what* grokking networks compute [Nanda et al., 2023, Chughtai et al., 2023, Gromov, 2023], and a recurring observation is that the effective rank (ER) of their weight matrices falls as test accuracy jumps [Liu et al., 2023b, DeMoss et al., 2025, Lei and Xu, 2026]. This has led to proposals that a pressure toward low-rank, small-norm or efficient circuits causes grokking [Nanda et al., 2023, Liu et al., 2023a, Varma et al., 2023]. We ask two questions. **(Q1)** Does this rank compression cause grokking, or follow it? **(Q2)** Does the learned solution lie in the subspace spanned by a few leading directions of each weight matrix?

*Corresponding author: shlok.mehendale@lossfunk.com

Preprint.

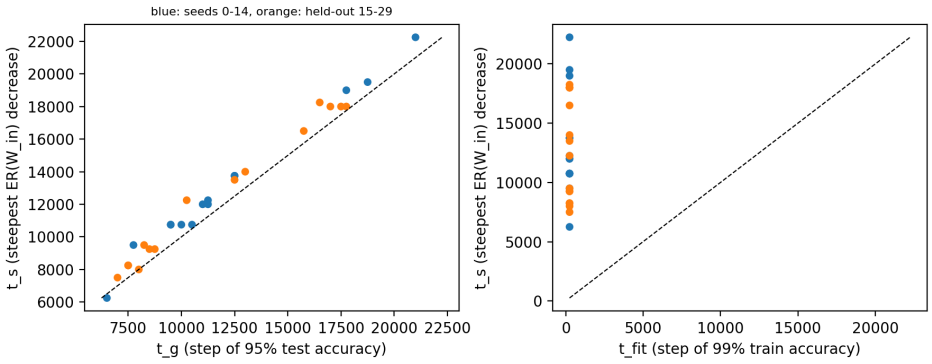
The usual evidence cannot settle these questions, for three reasons. First, rank compression is usually observed after two events, memorization (the first step at which training accuracy reaches 99%) and the rise in test accuracy (Figure 1a), so the timing alone cannot explain which of the two it follows. Second, timing without interventions cannot show cause. Lastly, the compactness of a solution is naturally measured by truncating each weight matrix to lower and lower rank until test accuracy drops. Keeping a weight matrix's k strongest singular directions gives the closest rank-k matrix, which acts only within their span. The span of the fewest top directions that leave the test accuracy unchanged, which we call the *sufficient subspace*, is then taken as the subspace the solution needs, analogous to the intrinsic dimension of Li et al. [2018]. Accuracy, however, only checks which output is largest. The margin between the correct-class output and the largest other output sets how far the outputs can move before a prediction changes. As margins grow, more directions can be deleted before any prediction changes, so the count can fall even though the network's outputs still depend on those directions.

We address these questions with geometric metrics (effective rank, stable rank, sparsity, inverse participation ratio) and multiple interventions (truncation of weight matrices to their top, bottom or random singular directions, complement interventions, transplant patching with orthogonal alignment, a nuclear-norm penalty, rank caps compared with factorized controls, and optimizer swaps). We study modular addition $(a+b) \bmod 97$ in two architectures with multiple random initializations. In the transformer, we analyze the embedding W_E, the unembedding W_U, and the input and output matrices W_{in} and W_{out} of its MLP block. We call the first step at which test accuracy reaches 95% the *grokking onset*. To separate memorization from the onset, we follow the transformer with a checkpoint every 250 steps and compare networks that grok with networks that memorize but never grok. To count directions without the margin, we judge each truncation by how much it changes the network's outputs. To test cause, we limit the rank of a weight matrix throughout training and compare each limit with a factorized full-rank control. We test the timing results on held-out transformer seeds and on two further tasks, $ab \bmod 97$ and $(a+b) \bmod 113$. Our results point to compression as a *signature* of grokking rather than its cause.

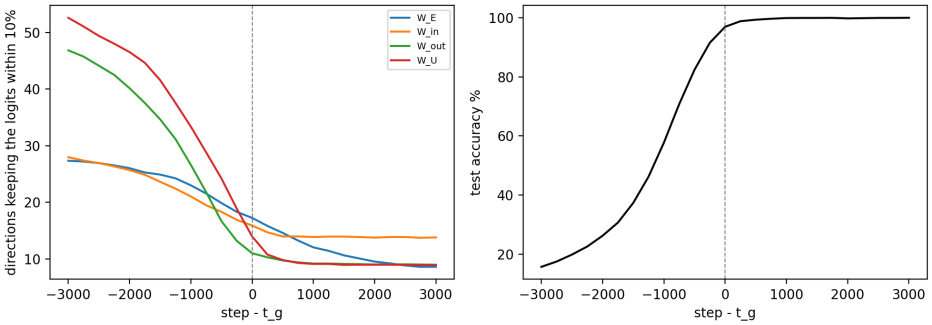
Contributions

- **Compression follows grokking, not memorization** (§6). Every transformer memorizes its training data within 250 steps, but the steepest fall in the effective rank of W_{in} comes about 900 steps after the grokking onset, whether the onset occurs at step 6,500 or 21,000. Networks that memorize without grokking lower this effective rank by 15–23%, against $60 \pm 5\%$ in those that grok. In the MLP, rank compression follows the onset in all 45 runs across three tasks. The timing replicates on 15 held-out seeds and, in the MLP, on two further tasks.
- **Subspace crystallization: the transformer's solution narrows to a low-dimensional subspace** (§5, §7). Before the transition, no weight matrix can be truncated below 27–53 directions without changing the network's outputs by more than 10%. After the transition, 8–13 directions suffice (Figure 1b). We call this narrowing of the subspace that the outputs depend on *subspace crystallization*. It is consistent with the removal of memorizing components described by Nanda et al. [2023]. We judge truncations by the outputs rather than only by the test accuracy. In the MLP trained with mean-squared error, truncation to 65 of 128 directions keeps the test accuracy, but keeping the outputs within 10% needs 122–127.
- **Imposing low rank before the transition does not speed up grokking** (§6.4). If rank compression drove grokking, removing directions in advance should help. Instead, against a factorized full-rank control, capping W_{in} at rank 32, 24 or 16, all above the 13 directions its solution needs, never lowers the mean onset, and capping W_E prevents grokking in 15 of 20 runs. Compressing the MLP's first layer with a nuclear-norm penalty before the onset also delays grokking, and comparing with the unfactorized network instead of the factorized control would wrongly suggest that low rank speeds grokking up, because the factorization changes the onset by itself.

We also report three secondary results. The late fall of effective rank in the MLP is set by AdamW's ϵ: once the square root of its second-moment estimate falls below ϵ, its updates reduce to gradient descent with momentum and learning rate lr/ϵ (§4). Different MLP seeds share their hidden representation up to an orthogonal rotation, but not with networks trained on another task or on random labels (§5). The Fourier inverse participation ratio (IPR), which measures how periodic the hidden units are [Doshi et al., 2024], rises before the onset but also in every network that never groks, so it cannot serve as a



(a) Step of the steepest effective-rank decrease of W_{in} against the grokking onset (left) and the step at which training accuracy reaches 99% (right). Blue: seeds 0–14. Orange: held-out seeds 15–29. Dashed line: identity.



(b) Number of directions of each weight matrix needed to keep the network's outputs within 10% (left) and test accuracy (right), with each seed aligned at its own onset t_g.

Figure 1: **Compression follows the grokking onset, and the directions needed to keep the outputs fall during the transition** (one-layer transformer, 29 grokked seeds). (a) Every seed fits its training data by step 250, yet compression follows the grokking onset across a threefold range of onset times. (b) The number of directions needed to keep the outputs within 10% falls during the rise in test accuracy and settles at 8–13 per matrix.

label-free early-warning signal (§6.5). Table 1 summarizes the evidence for the three contributions in both architectures.

2 Related Work

Grokking. Grokking was given its first mechanistic account by Nanda et al. [2023], who identified the learned algorithm as a sparse Fourier cosine decomposition, formed in phases of memorization, circuit formation and cleanup. Chughtai et al. [2023] extended this to general group composition, and Gromov [2023] established exact solvability with analytic weights. Proposed causes of the delay include circuit efficiency [Varma et al., 2023], the weight norm [Liu et al., 2023a], softmax collapse [Prieto et al., 2025] and a transition from lazy to rich training [Kumar et al., 2024]. Lei and Xu [2026] describe grokking as construction followed by compression, tracked through the network's Jacobian, consistent with our timing results. We add the separation of compression from memorization, negative controls and a replication on held-out seeds and tasks. Musat [2025] proves that, in the limit of small learning rate and weight decay, gradient descent after memorization minimizes the weight norm on the zero-loss manifold. Our MLP lies outside this regime: late in training, AdamW acts as gradient descent with a large learning rate $\text{lr}/\epsilon = 10^5$, and the weight norm grows (§4).

Complexity measures of grokking. Network complexity falls around the accuracy jump, whether measured by the linear mapping number [Liu et al., 2023b] or by the compressed description length of the weights [DeMoss et al., 2025]. We test whether this fall drives grokking or follows it, and show

Table 1: **Evidence for the three contributions in both architectures.** Effective rank is measured on W_1 (MLP) and W_{in} (transformer). Output counts are the fewest directions per matrix that change the network's outputs by at most 10%.

	MLP	Transformer
1. Compression follows grokking, not memorization		
Compression starts after test accuracy rises	45 of 45 runs	28 of 29 seeds
Fall in effective rank	10% if grokked	$60 \pm 5\%$ if grokked, 15–23% if not
2. Subspace crystallization: the transformer's solution contracts into a low-dimensional subspace as it groks		
Directions needed per matrix	Accuracy: from 110 to 65 of 128 Outputs: 122–127 of 128 (MSE)	Outputs: from 27–53 to 8–13, during the transition
3. Imposing low rank does not speed up grokking		
Rank caps against a factorized control (Table 4)	Caps delay the onset	W_{in}: onset unchanged or delayed W_E: 15 of 20 runs fail to grok

that a count of directions judged by accuracy can register compression that the network's outputs do not show (§5).

Low-rank bias and factorized controls. Gradient descent is biased toward low-rank solutions [Huh et al., 2023] and moves within a small subspace of parameter space [Gur-Ari et al., 2018]. Factorizing a matrix strengthens this bias [Arora et al., 2019] and accelerates optimization [Arora et al., 2018], so a rank cap written as a product $W = AB$ changes the optimization as well as the rank. We therefore compare each cap with a factorized full-rank control (§6.4).

Sparse circuits and universality. Merrill et al. [2023] describe grokking as a competition between a sparse generalizing subnetwork and a dense memorizing one, consistent with the few directions per matrix that the transformer's solution keeps. Networks trained on modular arithmetic can learn different algorithms [Zhong et al., 2023] or share one across architectures [Chughtai et al., 2023]. Our MLP seeds use different Fourier coordinates but share their hidden representation up to a rotation (§5). Appendix A discusses further work.

3 Setup

Architecture 1: mean-field MLP. A single-hidden-layer MLP with 128 hidden units and quadratic activations, $f(x) = \frac{1}{128} W_2\, \phi\big(W_1 x/\sqrt{194}\big)$ with $\phi(z) = z^2$ and input $x = [e_a; e_b] \in \mathbb{R}^{194}$, the concatenation of two one-hot vectors ($W_1 \in \mathbb{R}^{128\times 194}$, $W_2 \in \mathbb{R}^{97\times 128}$, i.i.d. $\mathcal{N}(0,1)$ initialization). The quadratic activation makes the Fourier computation explicit: for $x = \cos\theta$ it doubles frequencies, $x^2 = \frac{1}{2}(1 + \cos 2\theta)$, and applied to a sum of operand features it produces the products $\cos\theta_a \cos\theta_b$ from which $\cos(\theta_a + \theta_b)$ is read out. We train it with mean-squared error to one-hot targets, as in Gromov [2023], using full-batch AdamW (lr 10^{-3}, weight decay 10^{-2}, $\epsilon = 10^{-8}$) on a 35% train split for 10^5 steps, with 15 seeds. Five MLPs trained on randomly permuted labels serve as negative controls, and for the replication the MLP is also trained on $ab \bmod 97$ and $(a+b) \bmod 113$ (15 seeds each).

Architecture 2: one-layer transformer. Four attention heads, $d_{\text{model}} = 128$, $d_{\text{head}} = 32$, $d_{\text{mlp}} = 512$, ReLU MLP, no layer normalization and biases fixed at zero, following Nanda et al. [2023]. The inputs are $[a, b, =]$ and the prediction is read at the final position. We train it with cross-entropy, using full-batch AdamW (lr 10^{-3}, weight decay 1.0, $\beta = (0.9, 0.98)$) on a 30% train split for 25,000 steps, with 15 seeds. Fourteen of these seeds reach 95% test accuracy within this budget. The remaining seed, and three seeds each trained with weight decay 0.1 and 0.3, memorize the training set without generalizing and serve as negative controls. Seeds 15–29, trained only after the timing measure had been defined on seeds 0–14, form the held-out set of the replication, and all of them grok. For the time-resolved analyses, every run was repeated with a checkpoint every 250 steps. We analyze the embedding W_E, the unembedding W_U, the MLP projections W_{in} and W_{out}, and the attention projections W_Q, W_K, W_V, W_O (heads stacked).

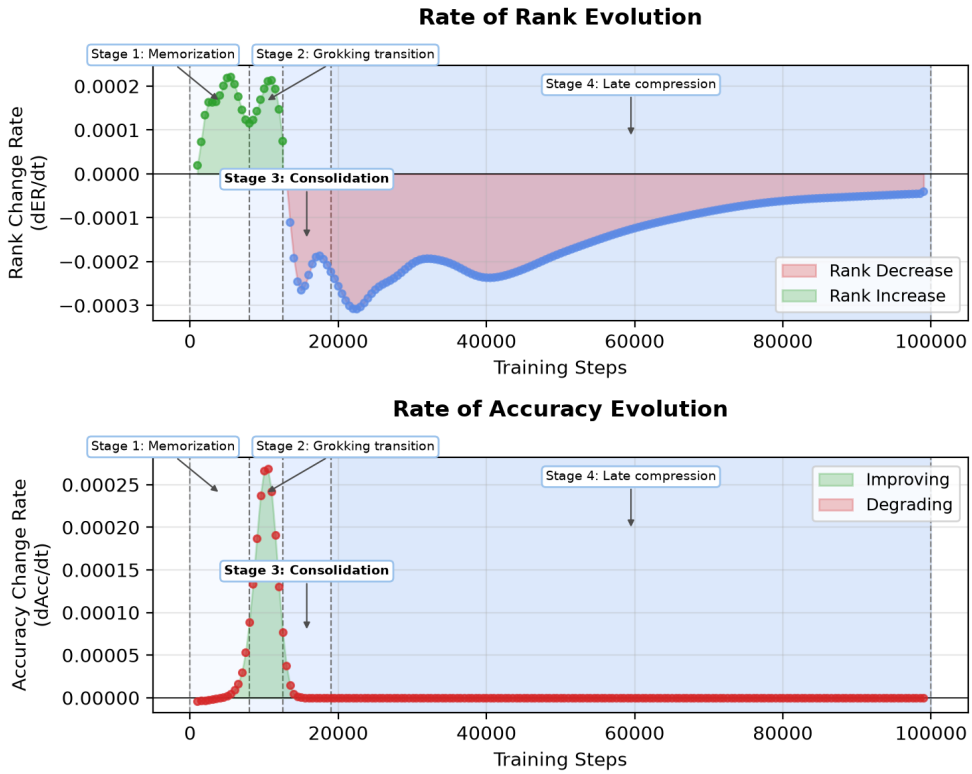


Figure 2: **The four training stages of the MLP** (seed 0): rate of change of the effective rank of W_1 (top) and of test accuracy (bottom), computed from the five-step moving average used for the steepest-decrease measure (Appendix E.1). Stage 2 starts at the first step with test accuracy above 5%, Stage 3 at t_g and Stage 4 at t_c. Effective rank grows during memorization and the grokking transition, turns negative at t_g and keeps falling through Stages 3 and 4.

Timing. The **grokking onset** t_g is the first logged step with test accuracy $\geq 95\%$, and the **memorization step** t_{fit} the first logged step with training accuracy $\geq 99\%$. A run counts as grokked if its final test accuracy is at least 95%. Throughout the paper, "before" and "after generalization" mean before and after t_g.

Effective rank. Our primary metric is the **effective rank** $\text{ER}(W) = \exp(H(\hat{\boldsymbol{\sigma}}))$ [Roy and Vetterli, 2007], the exponential of the entropy of the normalized singular values $\hat{\sigma}_i = \sigma_i / \sum_j \sigma_j$. A matrix whose k nonzero singular values are equal has $\text{ER} = k$. Appendix B defines the other tracked metrics (stable rank, sparsity, spectral IPR, weight and gradient norms), and Section 6.5 the Fourier IPR.

Code to reproduce all experiments is available at `https://github.com/Lossfunk/Causal-geometric-grokking`.

4 The Four-Stage Training Trajectory (MLP)

Figure 2 shows the four stages through the rates of change of effective rank and test accuracy. Numbers are means $\pm$ s.d. over 15 seeds, and Appendix Figure 3 shows all training metrics.

Stage 1 (Memorization): training accuracy rises to $97.8 \pm 0.4\%$ by step 8,000 while test accuracy stays below 5%. Test loss rises before it falls, a double-descent signature [Davies et al., 2023]. Effective rank rises slightly, from 114.6 for the random initial W_1 to about 116.

Stage 2 (Grokking transition): test accuracy rises from $37.0 \pm 8.0\%$ at step 10,000 to 95% at $t_g = 12{,}633 \pm 297$, and training accuracy reaches 100%. Effective rank rises only slightly (115.7 ± 0.4

Table 2: **Stage 4 tracks the effective learning rate.** Main seeds 0–2 are continued from step 20,000 to 10^5 with each optimizer, and entries are means. The effective learning rate is lr for gradient descent, $\text{lr}/(1-0.9)$ for SGD with momentum 0.9, and lr/ϵ for AdamW ($\text{lr} = 10^{-3}$). For $\epsilon = 10^{-9}$, lr/ϵ is an upper bound, since $\sqrt{\hat{v}}$ is comparable to ϵ. Weight decay is set so that the per-step shrinkage matches AdamW's. $d_{0.99}$: fewest directions that keep 99% of test accuracy.

Optimizer	Effective lr	$\|W_1\|_F$	$\sigma_1(W_1)$	ER(W_1)	$d_{0.99}$
Start (step 20,000)	–	726	111	115.1	65
AdamW, $\epsilon = 10^{-6}$	10^3	658	85	117.8	65
Gradient descent	10^3	662	85	117.8	65
AdamW, $\epsilon = 10^{-7}$	10^4	775	120	114.8	70
Gradient descent	10^4	783	126	114.6	72
AdamW, $\epsilon = 10^{-8}$ (main)	10^5	1,040	203	104.8	89
Gradient descent	10^5	833	148	108.3	87
SGD with momentum 0.9	10^5	1,093	216	103.8	91
AdamW, $\epsilon = 10^{-9}$	10^6	1,582	318	94.7	98
AdamW, $\epsilon = 10^{-8}$, no weight decay	10^5	1,257	261	97.8	94

Full sweep in Appendix C.2. Gradient descent at 10^6 diverges.

at step 10,000, 115.9 ± 0.6 at step 12,500), while the hidden units become periodic in their inputs, as measured by the Fourier IPR (§6.5), and the gradient norm peaks just after t_g.

Stage 3 (Consolidation): test accuracy is 100% and ER stays near its maximum until the compression onset $t_c = 19{,}000 \pm 3{,}279$, the first step at which it falls below 99% of its running maximum. Over the same interval the median margin grows from 0.105 ± 0.017 at step 12,500 to 0.470 ± 0.019 at step 20,000, and the directions of W_1 needed to keep 99% of test accuracy fall from 109.5 ± 3.2 to 64.9 ± 1.7, a consequence of the growing margin (§5). The stable rank of W_1 peaks around t_g and starts to fall.

Stage 4 (Late compression): ER falls to 104.0 ± 2.1 by step 10^5, and $99 \pm 2\%$ of its total decrease happens after t_g. The stable rank keeps falling, the weight norms grow throughout and the margin barely changes (0.599 to 0.611 between steps 50,000 and 10^5). Observing this stage requires training well past the point where test accuracy saturates.

Stage 4 is set by the effective learning rate. Stage 4 is growth of the leading singular directions of W_1 against weight decay: between steps 20,000 and 10^5 the largest singular value doubles (112 to 218) while the remaining 63 directions keep their size (Appendix C.1). The loss gradient that drives this growth is tiny (3.4×10^{-6} at step 20,000), points increasingly outward along the weights (its cosine with $-\theta$ rises from 0.19 to 0.89) and concentrates in the top 65 directions of W_1 (from 43% to 87% of its squared norm). For 87–100% of the coordinates, the square root of Adam's second-moment estimate lies below ϵ, so the update $m/(\sqrt{v} + \epsilon)$ is close to m/ϵ: AdamW acts as gradient descent with learning rate $\text{lr}/\epsilon = 10^5$ (Appendix C.2). Continuing the main runs from step 20,000 with other optimizers confirms that this effective learning rate, not Adam itself, governs Stage 4 (Table 2). At effective learning rates of 10^3 and 10^4, AdamW and gradient descent reach the same weights. At 10^5, gradient descent reproduces most of Stage 4 and SGD with momentum reproduces all of it (effective rank 108.3 and 103.8, against 104.8 under AdamW). At 10^3, weight decay shrinks the norm and effective rank stays at 117.8, while $\epsilon \leq 10^{-9}$ strengthens Stage 4.

Weight decay, loss and activation. The MLP (with MSE loss) groks without weight decay: from weight decay 0 to 0.1, t_g barely moves ($12{,}700 \pm 447$ to $13{,}400 \pm 418$), but the Stage-4 decrease in effective rank shrinks from $14.9 \pm 1.9\%$ to $0.5 \pm 0.4\%$ ($9.7 \pm 1.5\%$ at the default 10^{-2}), because weight decay restrains the growth of the leading directions. Trained with cross-entropy at learning rates 3×10^{-3} and 10^{-2}, the MLP groks in every seed ($t_g = 27{,}500 \pm 2{,}372$ and $10{,}100 \pm 962$) and shows Stage-4 decreases of $10.9 \pm 0.5\%$ and $37.9 \pm 1.9\%$. At lr 10^{-3} it groks in only one of five seeds within 10^5 steps. ReLU and GELU MLPs also grok, later than quadratic ones, with periodic

Table 3: d_{out} **and** $d_{0.99}$ **in the MLP** (directions of W_1 out of 128, on test inputs unless noted, mean $\pm$ s.d.). d_{out}: fewest directions that keep the outputs within 10%. $d_{0.99}$: fewest directions that keep 99% of test accuracy. Cross-entropy (CE) models use centered logits.

Model	Checkpoint	Outputs within 10%	99% of accuracy	Median margin
MSE (15 seeds)	step 0 (untrained)	116.4 ± 0.5	–	-0.004
	step 11,000 (before onset)	122.4 ± 0.5	119.3 ± 3.2	0.025
	step 20,000	122.3 ± 0.9	64.9 ± 1.7	0.470
	end (10^5)	127.3 ± 0.6	89.1 ± 2.3	0.611
MSE, random labels (5)	end, training inputs	123.0 ± 0.0	–	0.110
CE, lr 3×10^{-3} (5)	last before onset	121.4 ± 0.9	112.8 ± 5.5	3.8
	first after onset	119.0 ± 2.2	91.8 ± 15.7	5.9
	end	81.0 ± 1.6	50.6 ± 2.3	9.4
CE, lr 10^{-2} (5)	last before onset	120.2 ± 0.8	104.6 ± 4.1	4.2
	first after onset	119.8 ± 0.4	98.6 ± 4.2	4.9
	end	69.6 ± 2.2	48.4 ± 2.5	10.2

first-layer weights and Stage-4 decreases of 14–17% (Appendix B.2). With the same hyperparameters, the quadratic MLP groks 2.6 to 3.5 times sooner than ReLU and GELU MLPs, so the quadratic activation speeds up grokking in our setting but is not required for the MLP to grok (Appendix B.4).

5 Subspace Crystallization: Locating the Sufficient Subspace

This section defines the sufficient subspace and the output count d_{out}, and uses them to test whether the MLP crystallizes. The transformer's crystallization is shown in Section 7.

Definitions. We replace a weight matrix $W = U\Sigma V^\top$ by its top-k truncation $W_{[k]} = U_{:k}\Sigma_{:k}V_{:k}^\top$ and leave the rest of the network unchanged. $d_{0.99}(W)$ is the smallest k that keeps 99% of the test accuracy, and the span $\mathcal{C}$ of the top-$d_{0.99}$ left singular vectors is the *sufficient subspace*. Because accuracy only checks which output is largest, $d_{0.99}$ can fall as margins grow. $d_{\text{out}}(W)$ is the smallest k with $\|f_{[k]} - f\|_F \le 0.1\,\|f\|_F$ over the test inputs, where f stacks the network's outputs and, for cross-entropy models, each input's mean logit is removed, since softmax ignores it. *Subspace crystallization* is a narrowing of d_{out}. The leading directions carry the solution: truncation to the top $d_{0.99}$ directions keeps at least 99.26% of test accuracy, while random subspaces of equal rank keep 22.7% in the MLP and at most 12.5% in the transformer (Table 15).

In the MLP, crystallization is only apparent. With mean-squared error, crystallization is only apparent. The sufficient subspace of W_1 narrows right after the onset, from $d_{0.99} = 109.5 \pm 3.2$ at step 12,500 to 64.9 ± 1.7 at step 20,000, and widens again to 89.1 ± 2.3 by step 10^5 (Table 13). Over the narrowing, however, the median margin grows more than fourfold and d_{out} does not fall: it is 122.4 ± 0.5 before the onset, 122.3 ± 0.9 at step 20,000 and 127.3 ± 0.6 at the end (Table 3), above the 116.4 ± 0.5 of an untrained network and the 123.0 of a network trained on random labels. The same holds on both replication tasks (Appendix D.4). The narrowing of $d_{0.99}$ therefore measured how far the predictions were from flipping, not how many directions the function uses. Section 8 attributes the high d_{out} to the loss: mean-squared error to one-hot targets requires reproducing every Fourier component of the outputs. The later widening of the sufficient subspace comes from the growth of the leading singular directions, which makes the trailing directions matter more for the outputs (Appendix C.1). The complement $\mathcal{C}^\perp$ is unused for the predictions but is not inert: Gaussian noise added inside $\mathcal{C}^\perp$ lowers accuracy more than the same noise inside $\mathcal{C}$ (71% against 92%, Appendix D.3). With cross-entropy, which only requires the correct output to be largest, the MLP does crystallize, but late: $d_{\text{out}} \approx 120$ just before and just after the onset, and 81.0 ± 1.6 (learning rate 3×10^{-3}) and 69.6 ± 2.2 (10^{-2}) at the end of training (Table 3).

Equivalent solutions across random initializations. MLP seeds implement the same solution in different coordinates. Each seed spreads its first-layer Fourier power over most of the 48 frequencies (45.2 ± 0.6 hold 99%), with different strongest frequencies (Jaccard overlap 0.05 over 105 seed pairs), and the top-60 weight subspaces of different seeds overlap no more than random subspaces

(0.466 ± 0.027 against 0.487 ± 0.005 in mean cosine of the principal angles, Appendix D.5). Yet after an orthogonal Procrustes rotation fitted on training inputs [Schönemann, 1966], one seed's readout decodes another seed's post-activations with $95.8 \pm 2.4\%$ accuracy, against 1.0% without the rotation (Table 17). The rotation does not rescue a random network, a network trained on random labels or a network grokked on $ab \bmod 97$ (0.6–1.4%). Each seed thus implements the same task-relevant representation in a different orientation, consistent with Schiffman [2026], who find invariant algorithmic cores in independently trained transformers.

6 Timing and Cause of Compression

We test whether compression drives grokking with three kinds of evidence: its timing across seeds, how much the effective rank decreases in networks that memorize without grokking, and interventions that impose low rank from initialization.

6.1 Compression Follows the Onset, Not Memorization

The transformer separates memorization from the transition, and compression follows the transition. Every one of the 29 grokked seeds (seeds 0–14 and the held-out seeds 15–29) fits its training set within the first 250 steps, while their onsets range from 6,500 to 21,000 steps. With weight decay 1.0, AdamW multiplies the weights by 0.999 per step, so by step 4,000 the initial weights have decayed to 2% of their size. Because random weights spread over all directions, this decay alone lowers the effective rank of W_{in}, to a plateau at about 60% of its initial value, independently of grokking (Figure 9). We therefore measure compression only after step 4,000. After this decay, a second, sharp drop follows. The step t_s at which this second drop is steepest follows the onset almost exactly (Pearson correlation $r = 0.99$, slope 1.02, 95% confidence interval 0.98–1.07), scattering by 506 steps around t_g against 4,234 around the memorization step (Figure 1a). If compression followed memorization, t_s would cluster at a fixed step.

Networks that memorize without grokking compress far less. From step 4,000 to the end of training, the ER of W_{in} falls by $60 \pm 5\%$ in grokked transformers but by only 15–23% in the seven that never grok, at a peak rate two to five times lower (Appendix E.2). In the MLP, whose onset and memorization time barely vary across seeds, the random-label networks reach 90% training accuracy without grokking, and their effective rank falls by $1.1 \pm 0.3\%$, against $10.5 \pm 1.8\%$ in grokked MLPs. Comparing seeds at the same training step shows that low rank is acquired at the transition itself (Appendix Table 21): before their own onset, grokked seeds have as many directions as the seed that never groks, and after it far fewer.

6.2 Compression Does Not Start Before the Rise in Test Accuracy

In the MLP, the compression onset follows the grokking onset in every seed (Appendix Table 18 and Figure 11). All 15 seeds have $t_c > t_g$, with mean lag $6{,}367 \pm 3{,}248$ steps (one-sided $t = 7.59$, $p = 1.3 \times 10^{-6}$), i.e. $50 \pm 25\%$ of t_g. Granger tests agree: changes in test accuracy predict changes in effective rank in all 15 seeds, and the reverse holds in at most 6 (Appendix E.1).

In the transformer, compression is part of the transition. We locate the start t_{on} of the second drop as the first step of the stretch leading up to t_s in which the decrease is at least 20% of its peak rate, using a centered five-point moving average that adds no delay (Appendix E.1). The drop starts $1{,}888 \pm 992$ steps before the onset, but after test accuracy has first reached 10% in 28 of 29 seeds (median 1,750 steps after), at $34 \pm 22\%$ test accuracy. It is steepest 905 ± 506 steps after the onset (27 of 29 seeds), and 63% of it happens after the onset. In both architectures, then, compression does not start before test accuracy begins to rise. In the MLP it starts after the transition, and in the transformer during it.

6.3 Replication on Held-Out Tasks

We tested the timing results on seeds and tasks not used to define the measures, and report every test (Table 22). On the 15 held-out transformer seeds (H1), the steepest decrease came 850 ± 524 steps after the onset (14 of 15 seeds, the fifteenth at the onset, one-sided $t = 6.28$, $p = 1.0 \times 10^{-5}$), close to the 964 ± 499 steps of seeds 0–14, on which the measure had been defined. On $ab \bmod 97$ and $(a{+}b) \bmod 113$, the MLP's compression onset (H2) and steepest decrease (H3) followed the

Table 4: **Rank caps in the transformer** ($W = AB$ with inner dimension r, 5 seeds per arm, t_g averaged over the seeds that grok, mean $\pm$ s.d.). $r = 128$ is the factorized control without a rank limit.

Matrix	r	Grokked	t_g	Final test accuracy (%)
None (unfactorized)	–	5/5	$10{,}500 \pm 1{,}777$	100.0 ± 0.0
W_{in}	128	5/5	$14{,}100 \pm 4{,}095$	99.6 ± 0.3
	32	5/5	$14{,}350 \pm 5{,}730$	98.4 ± 1.6
	24	4/5	$16{,}312 \pm 3{,}332$	95.5 ± 7.2
	16	3/5	$16{,}417 \pm 6{,}682$	94.5 ± 7.7
	8	1/5	$12{,}500$	50.6 ± 39.7
W_E	128	5/5	$18{,}050 \pm 5{,}917$	100.0 ± 0.0
	32	1/5	$15{,}750$	31.8 ± 39.1
	24	4/5	$16{,}312 \pm 2{,}633$	83.1 ± 37.4
	16	0/5	–	11.9 ± 9.5
	8	0/5	–	13.7 ± 11.8

grokking onset in all 30 runs. The narrowing and re-widening of $d_{0.99}$ (H4, H5) and the effect of the factorization (H6) also replicated. With a Bonferroni correction over the nine t-tests, every test except H4 on $(a+b) \bmod 113$ ($p = 0.013$) remains significant.

6.4 Forcing Low Rank from Initialization

Forcing low rank before the transition does not bring grokking forward (Table 4, Table 25). We write a weight matrix as $W = AB$ with inner dimension r, initialized so that the entries of AB have the variance of the original initialization, and compare each rank cap with a factorized control whose inner dimension is at full rank. The control matters because the factorization alone changes the grokking time.

Transformer. The transformer's function needs only about 13 directions of W_{in} and 8 of W_E, so caps above these numbers leave the solution's capacity intact. Capping W_{in} at 32 directions leaves the onset where the factorized control has it ($14{,}350 \pm 5{,}730$ against $14{,}100 \pm 4{,}095$ steps), capping it at 24 or 16 delays it and makes it less reliable (four and three of five seeds grok), and capping it at 8, below the 13 directions the solution uses, prevents it in four of five seeds. Capping W_E is worse: five of twenty capped runs grok. The four that grok at rank 24 do so at $16{,}312 \pm 2{,}633$ steps, against $18{,}050 \pm 5{,}917$ for the control. Whatever the cap, the capped matrix ends at an effective rank of 8–12, like its factorized control.

MLP. In the MLP the factorization alone accelerates grokking seven- to eight-fold ($t_g = 1{,}667 \pm 58$ for $W_1 = AB$ at full rank against $12{,}483 \pm 340$ for an unfactorized W_1, logged every 50 steps, and 7.2 and 8.4 times on the two replication tasks), whereas it slows the transformer ($14{,}100 \pm 4{,}095$ against $10{,}500 \pm 1{,}777$ steps for W_{in}, Table 4). The factorization's own effect is consistent with two known mechanisms. For gradient descent on linear networks, Arora et al. [2018, §5] show that factorizing a matrix preconditions the updates of the product, combining an adaptive learning rate with a form of momentum, which can accelerate convergence. Weight decay on A and B penalizes the nuclear norm of W, since $\|A\|_F^2 + \|B\|_F^2 \geq 2\|AB\|_*$. This penalty is negligible at the MLP's weight decay of 10^{-2} but strong at the transformer's 1.0, consistent with the factorization speeding up the MLP and slowing the transformer. Against the factorized control, rank limits delay grokking ($2{,}450 \pm 100$ steps at $r = 64$, $4{,}133 \pm 29$ at $r = 32$), and models with $r \leq 16$ do not grok within 10^5 steps. Compared with the unfactorized W_1 alone, low rank would wrongly appear to accelerate grokking (Appendix E.4). Because the MLP's solution uses most of its directions, these delays say less about compression than the transformer caps do. A nuclear-norm penalty $\lambda\|W_1\|_*$ with $\lambda = 10^{-7}$ from step 0 compresses W_1 well before grokking (effective rank 78.5 ± 3.4 at step 11,500, against 115.9 ± 0.3) and delays grokking from $12{,}700 \pm 447$ to $16{,}200 \pm 447$ steps, and $\lambda \geq 3 \times 10^{-7}$ prevents learning (5 seeds per λ).

Table 5: **The transformer crystallizes during the transition** (29 grokked seeds, each aligned at its own onset t_g, with checkpoints every 250 steps). d_{out} of each matrix (test inputs, mean $\pm$ s.d.), test accuracy, and the restricted and excluded training losses of Nanda et al. [2023] (means).

Step $- t_g$	W_E	W_{in}	W_{out}	W_U	Test acc. (%)	Restricted	Excluded
$-3{,}000$	27.3 ± 2.8	28.0 ± 3.3	46.9 ± 7.6	52.6 ± 7.3	15.7	2.09	4.8
$-2{,}000$	26.0 ± 2.7	25.7 ± 3.3	40.2 ± 8.1	46.6 ± 7.9	26.2	1.37	8.1
$-1{,}000$	23.0 ± 3.1	21.0 ± 3.5	26.6 ± 7.4	33.4 ± 8.7	57.7	0.64	14.7
-500	19.9 ± 3.0	18.3 ± 2.9	16.7 ± 5.5	24.2 ± 8.0	82.4	0.26	16.6
0	17.2 ± 2.6	15.9 ± 2.3	11.0 ± 2.6	14.0 ± 4.9	96.9	0.12	16.2
$+1{,}000$	12.0 ± 2.2	13.9 ± 1.8	9.2 ± 1.9	9.1 ± 1.9	99.9	0.03	14.0
$+3{,}000$	8.6 ± 2.0	13.8 ± 1.7	9.0 ± 1.8	8.9 ± 1.7	100.0	0.12	18.4

6.5 IPR Is Not a Specific Early-Warning Signal

A hidden unit of the generalizing solution reads its input as a single cosine wave over the residues. The **Fourier IPR** measures how close the units are to this. Following Doshi et al. [2024], with $\hat{w}_k(f)$ the magnitude of the discrete Fourier transform of unit k's weights, normalized to unit ℓ_2 norm,

$$\text{IPR}_{\mathcal{F}}(W_1) = \frac{1}{128} \sum_{k=1}^{128} \sum_{f=0}^{p-1} \hat{w}_k(f)^4. \tag{1}$$

We use the weights on operand a only, where Doshi et al. [2024] average over each unit's input and output weights. A unit tuned to a single frequency has IPR $1/2$, and unstructured weights give about 0.02–0.03. For the transformer we compute it for the embedding, over the 97 tokens. Doshi et al. [2024] use it to separate generalizing from memorizing neurons and show that it rises during grokking. We test whether this rise can serve as a label-free early-warning signal of the onset.

IPR rises before the grokking onset, but also in every network that never groks, so it tracks the progress of training rather than an impending onset (Table 29). In the MLP, the Fourier IPR crosses a 3σ threshold at the end of whatever baseline window is chosen and fires in all five random-label networks. In the transformer, it fires before t_g in nearly every grokked seed, but also in all seven runs that do not grok (Appendix E.5).

7 Subspace Crystallization in the Transformer

Time-resolved results are means $\pm$ s.d. over the 29 grokked seeds, and final-checkpoint results over the 14 grokked seeds of 0–14.

The transformer crystallizes during the transition. With a checkpoint every 250 steps and each seed aligned at its own onset, d_{out} narrows during the rise in test accuracy (Figure 1b, Table 5). Three thousand steps before the onset, at 16% test accuracy, d_{out} is 27–53 per matrix. At the onset it is 11–17, and 3,000 steps later 9–14. For W_{in}, W_{out} and W_U it is half-way down 740–960 steps before the onset, in 28 or 29 of 29 seeds, at about 60% test accuracy. W_E narrows at the onset and keeps narrowing afterwards. The narrowing coincides with the cleanup phase of Nanda et al. [2023]. Their restricted loss, which keeps only the key-frequency components of the logits, is half-way down $1{,}966 \pm 647$ steps before the onset, as the generalizing circuit forms. Their excluded loss, which removes these components, is half-way up $1{,}026 \pm 646$ steps before the onset, as the memorizing components are removed, at the same time as d_{out} narrows. d_{out} adds that the cleanup reduces the directions the outputs depend on, not only those the predictions depend on, and the MLP trained with mean-squared error shows no equivalent (§5). Transformers that never grok narrow only slowly and stay two to nine times wider (Appendix F.2).

The final solution uses few directions. At the final checkpoint, every analyzed matrix except W_V is far below full rank (Table 30). $d_{0.99}$ is 7.7 ± 1.5 of 97 for W_E, 8.7 ± 1.5 of 128 for W_{out} and 13.0 ± 1.3 of 128 for W_{in}, and the MLP post-activations need 17.2 ± 3.3 of 512 directions (3.4%, truncating the mean-centered activation matrix). These counts are not inflated by the margin: d_{out} is 8.1–13.0 for W_E, W_{in}, W_{out} and W_U.

Runs that never grok keep high-rank matrices. The seven runs that memorize without generalizing have far higher-rank MLP and unembedding matrices. Counting the singular directions that carry 99% of $\sum_i \sigma_i^2$, they keep 68–84 for W_{in} (grokked: 18.1 ± 3.8), 61–79 for W_{out} (11.8 ± 2.5) and 62–73 for W_U (23.8 ± 2.8) (Table 33). With the comparison at matched steps in §6.1 (Table 21), this shows that low rank in these matrices is acquired at the transition, not by training to zero loss under weight decay.

8 Why the Transformer Crystallizes Further Than the MLP

We attribute this gap to the architecture. The transformer's embedding uses about four frequencies: 4.2 ± 0.7 of them hold 90% of its Fourier power. Each frequency requires two directions, a cosine and a sine, so this predicts about 8 directions, close to $d_{0.99}$ of W_E, W_{out} and W_U (7.7–8.8, Table 31). The directions attention writes to nearly coincide with those W_{in} reads from (mean cosine of the principal angles 0.923, against 0.248 for random subspaces), and equally so in runs that never grok (Appendix F). The MLP block therefore receives an input already confined to a few directions, and the directions its outputs do not need are removed during the transition, together with the memorizing components (§7). The MLP's first layer instead reads both one-hot operands directly, so each frequency costs it four directions, and it spreads its weights over most of the 48 frequencies (§5). No MLP we trained crystallizes during the transition.

Other differences between the models, i.e. loss and the weight decay do not account for the gap. The loss explains only why the cross-entropy MLP ends with fewer representative directions than the mean-squared-error MLP. A one-hot target has equal power at all 97 Fourier components, so for a network that matches its targets, omitting one frequency changes the outputs by $\sqrt{2/97} \approx 14\%$, more than d_{out} allows. CE constrains only the ordering of the logits, and the CE MLP ends at d_{out} = 70–81, against 122–127 with MSE. Raising its weight decay to the transformer's per-step shrinkage of 10^{-3} speeds up grokking, from step $10{,}100$ to $2{,}300$, but does not make it crystallize: d_{out} is about 120 just after the onset and ends at 118.4 instead of 69.6, and stronger decay prevents learning (Table 7). Only the CE MLP's d_{out} falls late, and its Stage-4 fall in effective rank is four times as large (37.9% against 9.7%, Table 6), which suggested that the late fall is Stage 4 itself (§4). Consistently, at every weight decay d_{out} ends close to the effective rank of W_1 (69.6 against 71.1, 100.4 against 98.7, 118.4 against 115.7). The cross-entropy MLP's late narrowing is thus a different process from the transformer's crystallization.

9 Discussion

What our measurements imply for rank analyses. Our results suggest four practices for measuring how many directions a network uses. First, count directions with a criterion that is independent of the margin, such as d_{out}, and compare the count with that of untrained or memorizing networks: in our MLP, $d_{0.99}$ reported a narrowing that the outputs did not undergo. Second, report the training step with every count: the MLP's $d_{0.99}$ grows from 65 to 89 after generalization, so it is not a fixed property of the trained model. Third, compare any intervention that imposes low rank with a control of the same parameterization: against the unfactorized MLP, rank caps would wrongly appear to accelerate grokking. Fourth, read weight-spectrum statistics such as IPR or effective rank as signals of grokking only next to negative controls, since they also move in networks that never grok. In the transformer, a few leading singular directions of each matrix carry the solution and truncating to them keeps the test accuracy, while random truncations of equal rank do not (Table 15, subject to the assumptions in Appendix D.3). In the MLP, one seed's hidden representation transfers to another after an orthogonal alignment fitted on its few thousand training inputs, although their weight subspaces overlap no more than random ones (§5).

Limitations. Our tasks are all modular arithmetic on cyclic groups, so a non-abelian group [Chughtai et al., 2023] would test the timing results more broadly. Whether language or vision models have a comparably clean sufficient subspace is a natural next question. Within our setting, the two architectures differ in more than their structure: they are also trained with different losses and for different lengths. We varied weight decay separately and found that it does not explain the difference in crystallization (Appendix B.3), but the cross-entropy MLPs need larger learning rates than the

mean-squared-error ones to grok within 10^5 steps (one of five seeds groks at 10^{-3}). Finally, the Granger tests capture only temporal order, so our causal conclusions rest on the rank interventions.

10 Conclusion

We asked whether rank compression causes grokking (Q1) and whether the learned solution comes to lie in a few directions of each weight matrix (Q2). For Q1, compression follows the grokking onset rather than memorization: it tracks the onset across a threefold range of onset times while the training set is memorized within 250 steps, it does not start before test accuracy rises, and forcing low rank from initialization does not bring grokking forward. Compression is therefore a signature of grokking rather than its cause. For Q2, the answer depends on how directions are counted. Judged by the outputs, the transformer crystallizes during the transition, to 8–13 directions per matrix, at the same time as its memorizing components are removed. The MLP trained with mean-squared error does not: its $d_{0.99}$ narrows only because its margins grow, and its late compression tracks the optimizer's effective learning rate. Counting directions by the outputs rather than by accuracy, and comparing any imposed low rank with a control of the same parameterization, is what makes these distinctions visible.

AI Use Statement

We used generative AI tools to retrieve and discover relevant literature and to draft and polish portions of the manuscript where appropriate. Any AI outputs were heavily edited by the authors. We did not use generative AI to formulate or refine the research hypotheses, develop the methodology, design the experiments, or interpret the results. The authors take full responsibility for the final content of this work, including its text, claims, code and data.

References

Roman Abramov, Felix Steinbauer, and Gjergji Kasneci. Grokking in the wild: Data augmentation for real-world multi-hop reasoning with transformers. In *Proceedings of the 42nd International Conference on Machine Learning*, Proceedings of Machine Learning Research, 2025. URL https://proceedings.mlr.press/v267/abramov25a.html.

Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the Optimization of Deep Networks: Implicit Acceleration by Overparameterization. In *International Conference on Machine Learning*, 2018. URL https://mlanthology.org/icml/2018/arora2018icml-optimization/.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems*, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/c0c783b5fc0d7d808f1d14a6e9c8280d-Paper.pdf.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Bilal Chughtai, Lawrence Chan, and Neel Nanda. A toy model of universality: Reverse engineering how networks learn group operations. In *Proceedings of the 40th International Conference on Machine Learning*, 2023. URL https://proceedings.mlr.press/v202/chughtai23a.html.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. 2023. URL https://arxiv.org/abs/2309.08600.

Xander Davies, Lauro Langosco, and David Krueger. Unifying grokking and double descent, 2023. URL https://arxiv.org/abs/2303.06173.

Branton DeMoss, Silvia Sapora, Jakob Foerster, Nick Hawes, and Ingmar Posner. The complexity dynamics of grokking. *Physica D: Nonlinear Phenomena*, 2025. URL https://www.sciencedirect.com/science/article/pii/S0167278925003367.

Darshil Doshi, Aritra Das, Tianyu He, and Andrey Gromov. To grok or not to grok: Disentangling generalization and memorization on corrupted algorithmic datasets. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.13061.

C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 1969.

Andrey Gromov. Grokking modular arithmetic. *arXiv:2301.02679*, 2023. URL https://arxiv.org/abs/2301.02679.

Guy Gur-Ari, Daniel A. Roberts, and Ethan Dyer. Gradient descent happens in a tiny subspace. *arXiv preprint arXiv:1812.04754*, 2018. URL https://arxiv.org/abs/1812.04754.

Minyoung Huh, Hossein Mobahi, Richard Zhang, Brian Cheung, Pulkit Agrawal, and Phillip Isola. The low-rank simplicity bias in deep networks. *Transactions on Machine Learning Research*, 2023. URL https://arxiv.org/abs/2103.10427.

Ahmed Imtiaz Humayun, Randall Balestriero, and Richard Baraniuk. Deep networks always grok and here is why. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. URL https://proceedings.mlr.press/v235/humayun24a.html.

Tanishq Kumar, Blake Bordelon, Samuel J. Gershman, and Cengiz Pehlevan. Grokking as the transition from lazy to rich training dynamics. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=vt5mnLVIVo.

Ruian Lei and Zenglin Xu. Geometric compression in grokking: The three-stage modular dynamics of transformers, 2026. URL https://openreview.net/forum?id=27fc8hXB5N.

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1804.08838.

Ziming Liu, Eric J. Michaud, and Max Tegmark. Omnigrok: Grokking beyond algorithmic data. In *International Conference on Learning Representations*, 2023a. URL https://arxiv.org/abs/2210.01117.

Ziming Liu, Ziqian Zhong, and Max Tegmark. Grokking as simplification: A nonlinear complexity perspective. In *UniReps: the First Workshop on Unifying Representations in Neural Models*, 2023b.

Kaifeng Lyu, Jikai Jin, Zhiyuan Li, Simon Shaolei Du, Jason D. Lee, and Wei Hu. Dichotomy of early and late phase implicit biases can provably induce grokking. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=XsHqr9dEGH.

Neil Mallinar, Daniel Beaglehole, Libin Zhu, Adityanarayanan Radhakrishnan, Parthe Pandit, and Mikhail Belkin. Emergence in non-neural models: Grokking modular arithmetic via average gradient outer product. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025. URL https://arxiv.org/abs/2407.20199.

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

Mohamad Amin Mohamadi, Zhiyuan Li, Lei Wu, and Danica Sutherland. Grokking modular arithmetic can be explained by margin maximization. In *NeurIPS 2023 Workshop on Mathematics of Modern Machine Learning*, 2023. URL https://openreview.net/forum?id=QPMfCLnIqf.

Depen Morwani, Benjamin L. Edelman, Costin-Andrei Oncescu, Rosie Zhao, and Sham M. Kakade. Feature Emergence via Margin Maximization: Case Studies in Algebraic Tasks. In *International Conference on Learning Representations*, 2024. URL https://mlanthology.org/iclr/2024/morwani2024iclr-feature/.

Tiberiu Musat. The geometry of grokking: Norm minimization on the zero-loss manifold. *arXiv:2511.01938*, 2025. URL https://arxiv.org/abs/2511.01938.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2301.05217.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv:2201.02177*, 2022. URL https://arxiv.org/abs/2201.02177.

Lucas Prieto, Melih Barsbey, Pedro A. M. Mediano, and Tolga Birdal. Grokking at the edge of numerical stability. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=TvfkSyHZRA.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, 2017. URL https://arxiv.org/abs/1706.05806.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *15th European Signal Processing Conference (EUSIPCO)*, 2007.

Joshua S. Schiffman. Transformers converge to invariant algorithmic cores. In *Mechanistic Interpretability Workshop at ICML 2026*, 2026. URL https://openreview.net/forum?id=y0uis4ccS6.

Peter H. Schönemann. A generalized solution of the orthogonal Procrustes problem. *Psychometrika*, 1966.

Vikrant Varma, Rohin Shah, Zachary Kenton, János Kramár, and Ramana Kumar. Explaining grokking through circuit efficiency, 2023. URL https://arxiv.org/abs/2309.02390.

Thomas Walker, Ahmed Imtiaz Humayun, Randall Balestriero, and Richard Baraniuk. GrokAlign: Geometric characterisation and acceleration of grokking, 2025. URL https://arxiv.org/abs/2506.12284.

David Yunis, Kumar Kshitij Patel, Samuel Wheeler, Pedro Henrique Pamplona Savarese, Gal Vardi, Karen Livescu, Michael Maire, and Matthew Walter. Rank minimization, alignment and weight decay in neural networks. In *High-dimensional Learning Dynamics 2024: The Emergence of Structure and Reasoning*, 2024. URL https://openreview.net/forum?id=u3sssLLu4y.

Ziqian Zhong, Ziming Liu, Max Tegmark, and Jacob Andreas. The clock and the pizza: Two stories in mechanistic explanation of neural networks. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2306.17844.

A Further Related Work

Grokking beyond modular arithmetic. Grokking also occurs for adversarial robustness [Humayun et al., 2024], in multi-hop reasoning [Abramov et al., 2025] and in Recursive Feature Machines [Mallinar et al., 2025]. Walker et al. [2025] remove the delay with spherical constraints on the residual stream. Our factorized control shows a related effect: factorizing W_1 at full rank shortens the delay seven- to eight-fold in the MLP (§6.4).

Theories of the late phase. Lyu et al. [2024] prove that homogeneous networks with large initialization and small weight decay move from kernel-like dynamics early in training to a max-margin solution late in training, and Mohamadi et al. [2023] show that a margin-maximization bias explains grokking in two-layer quadratic networks on modular arithmetic, with sample complexity $\tilde{O}(p^{5/3})$ against $O(p^2)$ for kernel methods. Their setting, an exponential-type loss with weight decay, differs from our mean-squared-error MLP, which groks at the same step without weight decay (Table 6). Morwani et al. [2024] show that Fourier features and irreducible representations emerge from implicit margin maximization, consistent with the Fourier structure our networks learn without any spectral penalty.

Weight geometry and representations. Li et al. [2018] train networks in low-dimensional random subspaces of parameter space, and Yunis et al. [2024] relate rank minimization, weight alignment and weight decay beyond grokking. Raghu et al. [2017] compare representations across networks with SVCCA, and our cross-seed results extend this comparison to grokking: the representations differ in raw coordinates and agree up to rotation. Sparse autoencoders [Bricken et al., 2023, Cunningham et al., 2023] decompose representations into features, a basis complementary to the singular directions used here.

B Experimental Details and Training Variants

B.1 Data, Training and Metrics

Data and training. The MLP uses a 35% train split (3,293 of the 9,409 pairs) and logs metrics every 500 steps. The transformer uses a 30% split (2,822 pairs) and logs every 250 steps. Its vocabulary comprises $p + 1 = 98$ tokens (the residues and the equals sign), so $W_E \in \mathbb{R}^{97\times128}$ (the residue rows), $W_U \in \mathbb{R}^{128\times98}$, $W_{\text{in}} \in \mathbb{R}^{128\times512}$, $W_{\text{out}} \in \mathbb{R}^{512\times128}$, $W_Q^h, W_K^h, W_V^h \in \mathbb{R}^{128\times32}$ per head and $W_O \in \mathbb{R}^{128\times128}$ (heads stacked). The transformer's higher weight decay (1.0 against 10^{-2}) follows Nanda et al. [2023]. MLP variants use seeds 0–4, rank-constrained MLPs seeds 0–2, and transformer rank caps seeds 0–4. All experiments were run on NVIDIA RTX A5000 GPUs.

Training objectives. The MLP minimizes the mean-squared error to one-hot targets, $\mathcal{L}_{\text{MSE}} = \frac{1}{p\,|\mathcal{D}_{\text{train}}|}\sum_{(a,b)} \|f(a,b) - e_y\|_2^2$ with $y = (a+b) \bmod p$, and the transformer the cross-entropy over its output logits $z \in \mathbb{R}^{p+1}$, $\mathcal{L}_{\text{CE}} = -\frac{1}{|\mathcal{D}_{\text{train}}|}\sum_{(a,b)} \log\big(\exp(z_y)/\sum_c \exp(z_c)\big)$.

Metrics. Stable rank is $\|W\|_F^2/\|W\|_2^2$. Hoyer sparsity is $(\sqrt{n} - \|w\|_1/\|w\|_2)/(\sqrt{n}-1)$ over the n entries w of W_1, and L0 sparsity the fraction of entries below 10^{-3} in magnitude. The spectral IPR $\sum_i \hat{\sigma}_i^4$ rises as spectral mass concentrates in a few directions. The gradient norm is the ℓ_2 norm of the full gradient. For the transformer's Fourier IPR, each of the 128 embedding dimensions is transformed over the 97 tokens.

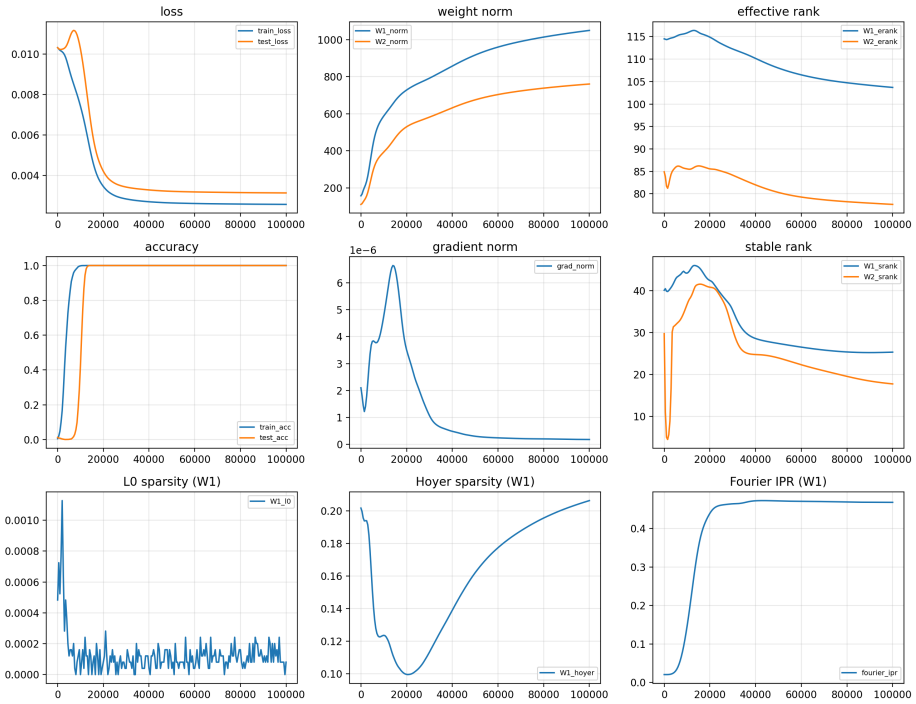


Figure 3: **Training dynamics of the MLP** (seed 0): loss, weight norms, effective rank, accuracy, gradient norm, stable rank, L0 and Hoyer sparsity of W_1, and Fourier IPR, against training step.

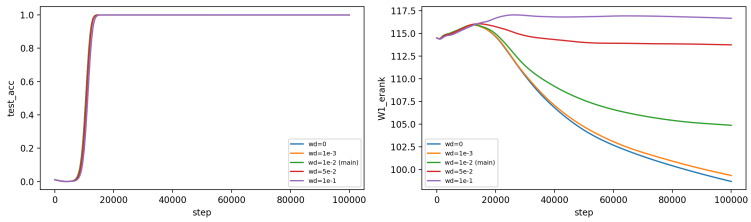
B.2 Training Variants of the MLP

Table 6 and Figure 4 report the variants summarized in Section 4.

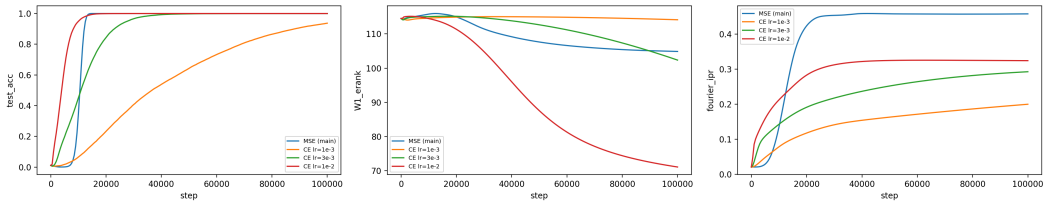
Table 6: **MLP training variants** (5 seeds each, mean $\pm$ s.d.). Main: mean-squared error, quadratic activation, weight decay (wd) 10^{-2}. CE: cross-entropy. Grokked: test accuracy $\geq 95\%$ within 10^5 steps. Stage-4 drop: relative decrease of effective rank from its peak after t_g to its final value. Fourier IPR: Section 6.5.

Variant	Grokked	t_g	Final ER	Stage-4 drop (%)	Fourier IPR
Main	5/5	$12{,}700 \pm 447$	104.9 ± 1.6	9.7 ± 1.5	0.457 ± 0.009
wd 0	5/5	$12{,}700 \pm 447$	98.7 ± 2.2	14.9 ± 1.9	0.452 ± 0.010
wd 10^{-3}	5/5	$12{,}700 \pm 447$	99.3 ± 2.1	14.4 ± 1.9	0.453 ± 0.008
wd 0.05	5/5	$12{,}800 \pm 274$	113.7 ± 0.8	2.2 ± 0.4	0.466 ± 0.006
wd 0.1	5/5	$13{,}400 \pm 418$	116.7 ± 1.7	0.5 ± 0.4	0.476 ± 0.004
CE, lr 10^{-3}	1/5	$94{,}500$ ($n = 1$)	114.1 ± 0.3	–	0.200 ± 0.008
CE, lr 3×10^{-3}	5/5	$27{,}500 \pm 2{,}372$	102.4 ± 0.6	10.9 ± 0.5	0.292 ± 0.009
CE, lr 10^{-2}	5/5	$10{,}100 \pm 962$	71.1 ± 2.2	37.9 ± 1.9	0.324 ± 0.011
ReLU	5/5	$44{,}900 \pm 4{,}615$	80.8 ± 0.8	14.0 ± 1.3	0.332 ± 0.028
GELU	5/5	$33{,}400 \pm 1{,}140$	79.2 ± 1.8	16.8 ± 1.3	0.474 ± 0.010

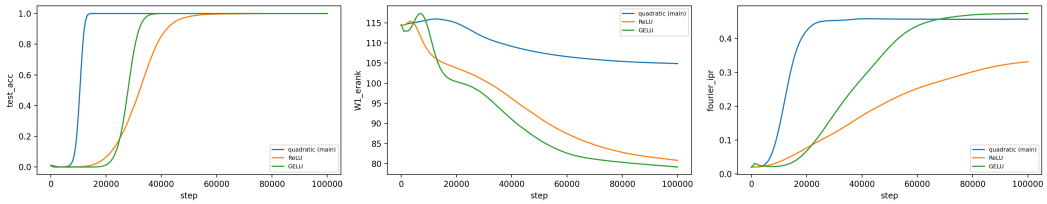
CE at lr 10^{-3} reaches $93.7 \pm 2.0\%$ final test accuracy, and its single grokked seed has a Stage-4 drop of 0.1%.



(a) Weight decay.



(b) Loss: mean-squared error vs. cross-entropy at three learning rates.



(c) Activation: quadratic, ReLU and GELU.

Figure 4: **Stage 4 across training variants** (MLP, 5 seeds each, mean curves of test accuracy, effective rank of W_1 and, in (b, c), Fourier IPR).

B.3 Weight Decay in the Cross-Entropy MLP

Table 7 trains the cross-entropy MLP at learning rate 10^{-2} with five weight decays λ (5 seeds each, 10^5 steps), including the runs of Table 6 at $\lambda = 10^{-2}$. With AdamW the weights shrink by $\text{lr} \cdot \lambda$ per step, so $\lambda = 0.1$ matches the transformer's per-step decay of 10^{-3}. d_{out} and $d_{0.99}$ are computed for W_1 on test inputs with centered logits, at the first saved checkpoint after t_g and at the end of training. At $\lambda \geq 0.3$ the weights decay to near zero, W_1 collapses to effective rank 1, and test accuracy stays at chance ($1.0 \pm 0.1\%$).

Table 7: **Weight decay in the cross-entropy MLP** (learning rate 10^{-2}, 5 seeds each, mean $\pm$ s.d.). d_{out} and $d_{0.99}$ of W_1 (of 128) on test inputs with centered logits, at the first checkpoint after t_g and at step 10^5.

λ	lr$\cdot\lambda$	Grokked	t_g	d_{out} after t_g	d_{out}, end	$d_{0.99}$, end	ER(W_1), end
0.01	10^{-4}	5/5	$10{,}100 \pm 962$	119.8 ± 0.4	69.6 ± 2.2	48.4 ± 2.5	71.1 ± 2.2
0.03	3×10^{-4}	5/5	$4{,}500 \pm 354$	120.2 ± 0.8	100.4 ± 2.3	61.6 ± 3.2	98.7 ± 2.6
0.1	10^{-3}	5/5	$2{,}300 \pm 274$	120.2 ± 0.8	118.4 ± 1.7	69.6 ± 2.6	115.7 ± 1.0
0.3	3×10^{-3}	0/5	–	–	–	–	1.0 ± 0.0
1.0	10^{-2}	0/5	–	–	–	–	1.0 ± 0.0

B.4 Role of the Nonlinearity

In the MLP, the quadratic activation turns sums of operand features into the products from which $\cos(\theta_a + \theta_b)$ is read out. ReLU and GELU do not do this exactly, yet MLPs with these activations also grok, at $t_g = 44{,}900 \pm 4{,}615$ and $33{,}400 \pm 1{,}140$ against $12{,}700 \pm 447$, with periodic first-layer weights (Fourier IPR 0.33 ± 0.03 and 0.47 ± 0.01, against 0.46 ± 0.01) and larger Stage-4 decreases in effective rank (Table 6). In the transformer, the ReLU of the MLP sublayer acts on inputs that

attention has written into the subspace the MLP reads (alignment 0.923, Table 34), and the embedding concentrates 95% of its Fourier power on five frequencies, so the MLP sublayer operates on an already low-dimensional, low-dimensional input.

C Stage 4: Weight Growth and the Optimizer

C.1 Why the Sufficient Subspace Widens

Table 8 and Figure 5 decompose the widening of the sufficient subspace for the main runs, and Table 9 repeats the measurement across weight decays. The top 65 directions of W_1 grow by half while the remaining 63 keep their size. The output change caused by removing the remaining directions, dominated by the cross term $2h_{\text{top}}h_{\text{rest}}$, roughly doubles while the margin grows by 30%, so truncation to the top 65 directions loses accuracy and $d_{0.99}$ widens. Stronger weight decay limits the growth of the top directions and, with it, both the fall in effective rank and the widening.

Table 8: **Why the sufficient subspace widens in Stage 4** (MLP, 15 seeds). $K = 65$ is the median $d_{0.99}$ at step 20,000. Size: $(\sum_i \sigma_i^2)^{1/2}$ over the top 65 or the remaining 63 singular directions of W_1. Output change: median ℓ_2 change of the 97 outputs when the directions beyond the top 65 are removed, split into the two terms of $(h_{\text{top}} + h_{\text{rest}})^2 - h_{\text{top}}^2$.

Step	20,000	30,000	50,000	100,000
$d_{0.99}$	64.9 ± 1.7	69.3 ± 1.7	81.5 ± 2.4	89.1 ± 2.3
ER(W_1)	114.4 ± 1.1	110.9 ± 1.5	107.0 ± 1.9	104.0 ± 2.1
$\|W_1\|_F$	731 ± 7	813 ± 14	937 ± 26	1059 ± 35
$\|W_2\|_F$	531 ± 5	597 ± 11	685 ± 21	763 ± 27
$\sigma_1(W_1)$	111.9 ± 3.4	132.2 ± 3.0	171.1 ± 10.4	217.8 ± 21.6
Size of top 65	659 ± 10	748 ± 18	878 ± 29	1002 ± 37
Size of remaining 63	316 ± 10	318 ± 10	327 ± 13	344 ± 18
Test acc., top-65 truncation (%)	99.1 ± 0.4	97.7 ± 0.9	84.5 ± 5.5	57.7 ± 9.5
Median correct-class output	0.604 ± 0.017	0.696 ± 0.010	0.722 ± 0.008	0.732 ± 0.007
Median margin	0.470 ± 0.019	0.570 ± 0.012	0.599 ± 0.009	0.611 ± 0.008
Output change, $2h_{\text{top}}h_{\text{rest}}$	0.611 ± 0.015	0.699 ± 0.014	0.860 ± 0.063	1.153 ± 0.128
Output change, h_{rest}^2	0.254 ± 0.019	0.336 ± 0.022	0.445 ± 0.029	0.533 ± 0.047

Table 9: **Weight decay controls the widening** (5 seeds each, step 20,000 $\rightarrow$ step 10^5, means). All runs reach 100% test accuracy.

Weight decay	$d_{0.99}$	ER(W_1)	$\|W_1\|_F$	$\sigma_1(W_1)$	Size of remaining 63
0	$65.8 \rightarrow 95.0$	$114.6 \rightarrow 98.7$	$731 \rightarrow 1{,}252$	$114 \rightarrow 251$	$320 \rightarrow 358$
10^{-3}	$65.4 \rightarrow 93.8$	$114.7 \rightarrow 99.3$	$731 \rightarrow 1{,}220$	$113 \rightarrow 249$	$320 \rightarrow 354$
10^{-2} (main)	$65.2 \rightarrow 88.6$	$115.0 \rightarrow 104.9$	$727 \rightarrow 1{,}043$	$111 \rightarrow 205$	$320 \rightarrow 347$
5×10^{-2}	$64.0 \rightarrow 73.0$	$115.7 \rightarrow 113.7$	$710 \rightarrow 825$	$101 \rightarrow 151$	$321 \rightarrow 342$
10^{-1}	$62.4 \rightarrow 63.0$	$116.7 \rightarrow 116.7$	$678 \rightarrow 748$	$92 \rightarrow 108$	$311 \rightarrow 346$

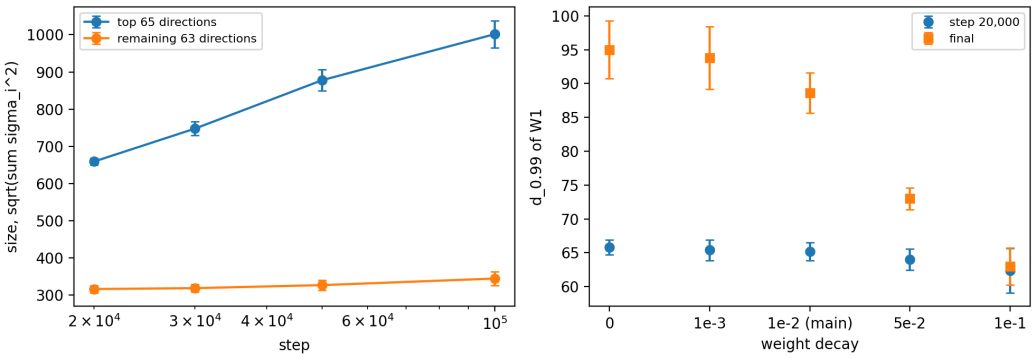


Figure 5: **Stage-4 widening.** Left: size of the top 65 and the remaining 63 singular directions of W_1 over training (15 seeds). Right: $d_{0.99}$ at step 20,000 and at the end of training for each weight decay (5 seeds).

C.2 Gradient and Adam State

We continue each main seed for 1,000 AdamW steps from the checkpoints at steps 20,000, 50,000 and 10^5 to rebuild the optimizer state, and then measure the full-batch gradient g and Adam's bias-corrected moments $\hat{m}, \hat{v}$ (Table 10). The gradient is small and becomes increasingly radial: its cosine with θ goes from -0.19 to -0.89, so $-g$ points outward and the loss would decrease if the weights grew. This radial component equals the Euler expression $\langle g, \theta \rangle = 3 \cdot \frac{2}{p|\mathcal{D}|} \sum (f - y) f$ for this 3-homogeneous model to all printed digits. Almost every coordinate has $\sqrt{\hat{v}} < \epsilon = 10^{-8}$, so the update $\hat{m}/(\sqrt{\hat{v}} + \epsilon)$ is close to $\hat{m}/\epsilon$: AdamW acts as gradient descent with momentum and learning rate $\mathrm{lr}/\epsilon = 10^5$, not as sign descent (the mean of $|\hat{m}|/(\sqrt{\hat{v}} + \epsilon)$ is 0.07–0.21, against 1 for sign descent). The radial part of this update exceeds the decay pull at step 20,000 and approaches it by the end, as the norm approaches an equilibrium. The gradient also concentrates in the top 65 directions of W_1, the directions that grow.

Table 10: **Gradient and Adam state during Stage 4** (MLP, 15 seeds, mean ± s.d.). Radial push of the update and decay pull $\lambda \|\theta\|^2$ are per unit learning rate.

	Step 20,000	Step 50,000	Step 10^5
$\|g\|$	3.4×10^{-6}	3.2×10^{-7}	1.9×10^{-7}
$\langle g, \theta \rangle$ (= Euler expression)	-6.0×10^{-4}	-2.5×10^{-4}	-2.2×10^{-4}
$\cos(g, \theta)$	-0.19 ± 0.01	-0.67 ± 0.07	-0.89 ± 0.06
Share of coordinates with $\sqrt{\hat{v}} < \epsilon$	0.870 ± 0.017	0.998 ± 0.001	0.9998 ± 0.0002
Median $\sqrt{\hat{v}}$	2.1×10^{-9}	7.3×10^{-10}	6.1×10^{-10}
Mean $\|\hat{m}\|/(\sqrt{\hat{v}} + \epsilon)$ (1 = sign descent)	0.21 ± 0.01	0.08 ± 0.01	0.07 ± 0.00
Radial push of the Adam update	$(2.0 \pm 0.1) \times 10^4$	$(2.0 \pm 0.2) \times 10^4$	$(1.9 \pm 0.1) \times 10^4$
Decay pull $\lambda \|\theta\|^2$	$(0.84 \pm 0.02) \times 10^4$	$(1.36 \pm 0.08) \times 10^4$	$(1.71 \pm 0.12) \times 10^4$
Share of $\|g_{W_1}\|^2$ in the top 65 directions	0.43 ± 0.03	0.79 ± 0.09	0.87 ± 0.03

C.3 Optimizer Swap

Table 11 continues main seeds 0–2 from step 20,000 to 10^5 with each optimizer. The weight decay is set so that the per-step shrinkage matches AdamW's $\mathrm{lr} \cdot \lambda = 10^{-5}$ ($\lambda = 10^{-5}/\mathrm{lr}$ for gradient descent, $10^{-6}/\mathrm{lr}$ for momentum 0.9, whose steady-state step is $\mathrm{lr}/(1 - 0.9)$). Runs with the same effective learning rate agree closely: SGD with momentum at lr 10 reproduces gradient descent at lr 100 to the printed digits, and AdamW with $\epsilon = 10^{-6}$ reproduces gradient descent at lr 10^3. Stage 4, the fall of effective rank with the growth of σ_1 and the widening of $d_{0.99}$, appears at effective learning rates of 10^5 and above and strengthens as ϵ decreases. Test accuracy stays at 100% in every run.

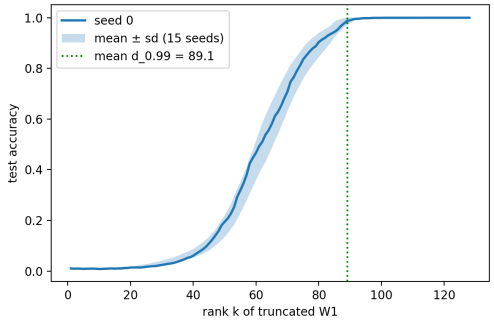
Table 11: **Continuations from step 20,000 to** 10^5 (main seeds 0–2, means). The effective learning rate is lr for gradient descent, lr/ϵ for AdamW and $\mathrm{lr}/(1-0.9)$ for SGD with momentum.

Optimizer	Effective lr	$\|W_1\|_F$	$\|W_2\|_F$	$\sigma_1(W_1)$	$\mathrm{ER}(W_1)$	$d_{0.99}$
At step 20,000	–	726	529	111	115.1	65
AdamW, $\epsilon = 10^{-10}$	10^7	1,793	718	346	95.7	99
AdamW, $\epsilon = 10^{-9}$	10^6	1,582	874	318	94.7	98
AdamW, $\epsilon = 10^{-8}$ (main)	10^5	1,040	748	203	104.8	89
AdamW, $\epsilon = 10^{-7}$	10^4	775	551	120	114.8	70
AdamW, $\epsilon = 10^{-6}$	10^3	658	469	85	117.8	65
AdamW, $\epsilon = 10^{-4}$	10	331	241	51	115.3	65
AdamW, $\epsilon = 10^{-8}$, no weight decay	10^5	1,257	917	261	97.8	94
SGD, lr 10^{-3} (same decay)	10^{-3}	326	238	50	115.1	65
Gradient descent	10^2	384	277	56	116.7	65
Gradient descent	10^3	662	472	85	117.8	65
Gradient descent	10^4	783	557	126	114.6	72
Gradient descent	10^5	833	913	148	108.3	87
Gradient descent	10^6			diverges		
SGD + momentum, lr 10	10^2	384	277	56	116.7	65
SGD + momentum, lr 10^2	10^3	662	472	85	117.8	65
SGD + momentum, lr 10^3	10^4	783	557	126	114.6	72
SGD + momentum, lr 10^4	10^5	1,093	775	216	103.8	91
SGD + momentum, lr 10^5	10^6	881	1,577	174	101.1	93

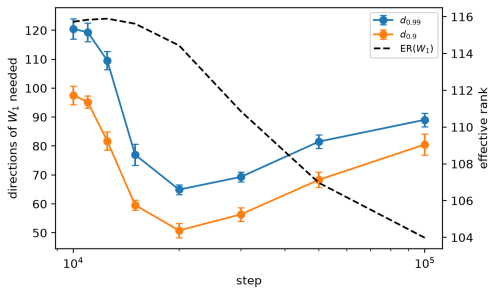
D The Sufficient Subspace, Output Counts and Cross-Seed Equivalence

D.1 The Sufficient Subspace over Training

Table 12 reports d_τ of W_1 at four thresholds at the end of training, which ranges from 80 to 93 directions for thresholds from 90% to 99.9%. Table 13 and Figure 6 show its evolution over training. Before step 15,000, d_τ is relative to each checkpoint's own test accuracy, which is below 100%. The number of frequencies holding 90% of the Fourier power of W_1 falls from 38.2 to 33.7 during Stage 4.



(a) Test accuracy of W_1 truncated to its top k directions (step 10^5).



(b) $d_{0.9}$ and $d_{0.99}$ (left axis) and effective rank (right axis) over training.

Figure 6: **The sufficient subspace of the MLP** (15 seeds). (a) Seed 0 and mean ± s.d. (b) $d_{0.9}$ and $d_{0.99}$ narrow sharply after the onset and widen again during Stage 4, while effective rank falls. d_{out} does not narrow (Table 3).

Table 12: d_τ of W_1 at four thresholds (step 10^5, 15 seeds).

Threshold τ (fraction of clean accuracy)	d_τ (mean $\pm$ s.d.)
90%	80.5 ± 3.7
95%	84.5 ± 3.2
99%	89.1 ± 2.3
99.9%	93.1 ± 2.7

Table 13: $d_{0.9}$, $d_{0.99}$, effective rank and frequency usage of W_1 over training (15 seeds).

Step	Test acc. (%)	$d_{0.9}$	$d_{0.99}$	ER(W_1)	Freq. for 90% power
10,000	37.0 ± 8.0	97.5 ± 3.2	120.5 ± 3.5	115.7 ± 0.4	40.8 ± 0.4
11,000	67.3 ± 8.9	95.2 ± 2.1	119.3 ± 3.2	115.8 ± 0.4	40.3 ± 0.7
12,500	95.7 ± 2.7	81.7 ± 3.2	109.5 ± 3.2	115.9 ± 0.6	39.7 ± 0.5
15,000	100.0 ± 0.0	59.6 ± 1.8	77.0 ± 3.6	115.6 ± 0.8	38.5 ± 0.6
20,000	100.0 ± 0.0	50.8 ± 2.5	64.9 ± 1.7	114.4 ± 1.1	38.2 ± 1.1
30,000	100.0 ± 0.0	56.4 ± 2.4	69.3 ± 1.7	110.9 ± 1.5	38.3 ± 0.9
50,000	100.0 ± 0.0	68.3 ± 2.6	81.5 ± 2.4	107.0 ± 1.9	36.1 ± 1.4
100,000	100.0 ± 0.0	80.5 ± 3.7	89.1 ± 2.3	104.0 ± 2.1	33.7 ± 1.9

D.2 Temporal Stability

We compute the mean cosine of the principal angles between the top-89 left and right singular subspaces of W_1 at step s and at the final checkpoint ($k = 89$, the median $d_{0.99}$), across all 15 seeds (Table 14). At initialization the overlap is at the random level. At t_g it is only modestly above it (0.837 against 0.773 in the hidden space), and it keeps rising for tens of thousands of steps. The final sufficient subspace is therefore assembled mostly during Stages 3 and 4, not at the accuracy transition.

Table 14: Overlap of the top-89 singular subspaces of W_1 at step s with those at step 10^5 (15 seeds). Hidden space: left singular vectors (the sufficient subspace $\mathcal{C}$). Input space: right singular vectors.

Step s	Hidden space ($\mathcal{C}$)	Input space
0	0.775 ± 0.003	0.608 ± 0.003
5,000	0.782 ± 0.003	0.623 ± 0.003
10,000	0.807 ± 0.005	0.677 ± 0.007
12,500	0.837 ± 0.007	0.723 ± 0.010
15,000	0.875 ± 0.010	0.773 ± 0.012
20,000	0.925 ± 0.013	0.843 ± 0.013
30,000	0.961 ± 0.013	0.910 ± 0.011
50,000	0.983 ± 0.012	0.967 ± 0.013
Random	0.773 ± 0.003	0.606 ± 0.003

D.3 Truncation Controls and Complement Interventions

Table 15: **Top-k, bottom-k and random-k truncations** at equal rank $k = d_{0.99}$ (per seed). Test accuracy in %, mean $\pm$ s.d. over 15 MLP seeds at step 10^5 and 14 grokked transformer seeds. Random controls average 10 draws. Chance is 1.03%.

Matrix	$k = d_{0.99}$	Top-k	Bottom-k	Random subset	Random subspace
MLP W_1	89.1 ± 2.3 / 128	99.26 ± 0.22	$28.26 \pm 6.18^*$	31.26 ± 4.37	22.74 ± 3.93
Transformer W_{in}	13.0 ± 1.3 / 128	99.80 ± 0.28	$5.50 \pm 1.46^\ddagger$	3.50 ± 0.54	12.49 ± 2.99
Transformer W_{out}	8.7 ± 1.5 / 128	99.95 ± 0.07	$5.50 \pm 1.46^\ddagger$	4.42 ± 0.95	8.00 ± 1.93
Transformer W_E	7.7 ± 1.5 / 97	99.74 ± 0.35	1.01 ± 0.06	1.33 ± 0.23	1.31 ± 0.13
Transformer W_U	8.8 ± 1.4 / 98	99.86 ± 0.27	1.07 ± 0.32	1.80 ± 0.65	6.02 ± 1.92

*Top-k and bottom-k share about 50 directions when $k > 64$. ‡Keeping only the bottom directions of W_{in} or of W_{out} effectively removes the MLP's contribution, hence the identical values.

Complement interventions. We intervene on the MLP's pre-activations $h = W_1 x/\sqrt{194}$, with $\mathcal{C}$ the sufficient subspace of each seed at step 10^5. Keeping only the $\mathcal{C}$-component gives $99.26 \pm 0.22\%$ test accuracy and keeping only the $\mathcal{C}^\perp$-component $0.15 \pm 0.14\%$. Adding Gaussian noise of norm $s\|h\|$ (per input) inside $\mathcal{C}$ or inside $\mathcal{C}^\perp$ gives $99.97 \pm 0.05\%$ and $99.95 \pm 0.07\%$ for $s = 0.1$, $92.42 \pm 4.22\%$ and $70.97 \pm 9.69\%$ for $s = 0.3$, $10.11 \pm 2.06\%$ and $7.43 \pm 1.32\%$ for $s = 1$, and $1.54 \pm 0.10\%$ and $1.61 \pm 0.09\%$ for $s = 3$. As a check of the pipeline, patching the clean hidden representation into a run on noise-corrupted inputs restores the clean output (100.00% in the MLP at corruption level 0.3, against 1.59% unpatched, and 99.92% in the transformer, against 18.52%), as it must in a network with one hidden layer.

Assumptions of the SVD probe. Equating high-variance singular directions with functional relevance assumes that those directions carry the computation rather than optimizer noise, representation drift or dense isotropic features. It also assumes linear, orthogonal structure, which can obscure sparse features stored in non-orthogonal superposition or on nonlinear manifolds. The controls of Table 15 test the first assumption in our setting: the high-variance singular directions carry the computation, while random subspaces of equal rank do not. In the MLP the distinction is weaker, because the sufficient subspace spans most of the directions.

D.4 Output Counts under Other Criteria

Table 16 gives the directions of W_1 needed under four output-based criteria: outputs within 10% (d_{out}) and 5%, per-input centered outputs within 10%, and per-input margins within 10%. Centering changes nothing for the MSE models. The margin criterion lies between accuracy and the full outputs: it narrows from 124 to 113 directions between steps 11,000 and 20,000 and widens again to 121, much less than $d_{0.99}$ (119 to 65 to 89). For the cross-entropy MLPs, every output-based criterion narrows only after the onset.

Table 16: **Output counts in the MLP** (directions of W_1 out of 128, mean over seeds, test inputs unless noted). d_{out} and Out05: outputs within 10% and 5% (centered for CE). Cen10: per-input centered outputs within 10%. Marg10: per-input margins within 10%.

Model	Checkpoint	$d_{0.99}$	d_{out}	Out05	Cen10	Marg10	Median margin
MSE, main	step 0	–	116.4	123.8	116.3	102.2	-0.004
	step 11,000	119.3	122.4	126.6	122.7	123.8	0.025
	step 20,000	64.9	122.3	125.7	122.4	113.1	0.470
	end	89.1	127.3	127.9	127.3	121.4	0.611
	end, training inputs	88.0	127.3	127.9	127.5	121.1	0.649
MSE, random labels	end, training inputs	–	123.0	127.0	123.2	116.6	0.110
	end, test inputs	–	121.4	126.0	121.6	108.0	-0.229
MSE, $ab \bmod 97$	last before t_g	115.9	122.1	126.3	122.3	117.5	0.052
	first after t_g	97.5	122.5	126.3	122.5	112.7	0.155
	step 20,000	64.7	121.7	125.1	121.7	112.4	0.481
	end	89.8	126.2	126.9	126.2	120.7	0.653
MSE, $(a+b) \bmod 113$	last before t_g	117.7	123.9	127.1	124.2	123.1	0.048
	first after t_g	79.4	124.3	127.1	124.4	115.3	0.232
	step 20,000	70.9	124.3	126.9	124.3	114.1	0.275
	end	85.5	127.4	127.9	127.4	121.5	0.501
CE, lr 3×10^{-3}	last before t_g	112.8	121.4	125.8	121.4	120.8	3.78
	first after t_g	91.8	119.0	125.4	119.0	112.6	5.86
	end	50.6	81.0	99.8	81.0	68.0	9.43
CE, lr 10^{-2}	last before t_g	104.6	120.2	125.8	120.2	117.8	4.22
	first after t_g	98.6	119.8	125.8	119.8	115.6	4.86
	end	48.4	69.6	71.8	69.6	65.8	10.20

D.5 Cross-Seed Equivalence

For a source model A and a receiving model B, we compute A's post-activations Z_A on A's test inputs and decode them with B's readout W_2^B, either directly or after the orthogonal map $R^\star = \arg\min_{R^\top R = I} \|Z_A^{\text{tr}} R - Z_B^{\text{tr}}\|_F$ fitted on A's training inputs [Schönemann, 1966], and measure accuracy against the true labels $(a+b) \bmod 97$ (Table 17). Rows with random-label models test specificity in one direction only: a readout trained on random labels cannot decode the true labels from any representation, so only transplants into a grokked readout are informative. The step-10,000 checkpoint of each seed (at least 99% training and $37.0 \pm 8.0\%$ test accuracy) already partly generalizes, so its rows do not isolate a memorizing representation.

Table 17: **Cross-seed transplants and linear CKA** (MLP, final checkpoints unless noted, mean $\pm$ s.d.). Transplant: test accuracy of model B's readout applied to model A's post-activations after the map (chance 1.03%). CKA on all p^2 inputs.

Model $A \to$ grokked model B	Map	Test accuracy (%)
Grokked, other seed	fitted orthogonal	95.82 ± 2.38
Grokked, other seed	none	1.03 ± 0.30
Grokked, other seed	random orthogonal	0.94 ± 0.27
Random initialization	fitted orthogonal	0.61 ± 0.12
Random-label network	fitted orthogonal	0.68 ± 0.07
Grokked on $ab \bmod 97$ (other task)	fitted orthogonal	1.42 ± 0.09
Same seed at step 10,000	none	36.18 ± 6.40
Same seed at step 10,000	fitted orthogonal	25.09 ± 5.77
Linear CKA between		
grokked seeds		0.096 ± 0.019
a grokked network and its initialization		0.089 ± 0.010
two initializations		0.142 ± 0.002

Subspace overlap. Figure 7 shows the Fourier power spectrum of W_1 for six seeds. Overlaps are mean cosines of the principal angles between the top-60 weight subspaces of W_1 (in the 194-dimensional input space), and between 10-dimensional Fourier subspaces built from each seed's five strongest frequencies (a cosine and a sine for each, with f and $p - f$ treated as one frequency),

over all 105 seed pairs, with random baselines simulated for the same dimensions. A few individual directions are shared: per seed pair, 5.78 ± 1.97 of the top-60 directions have a counterpart with cosine above 0.9, against 1.38 expected for random subspaces.

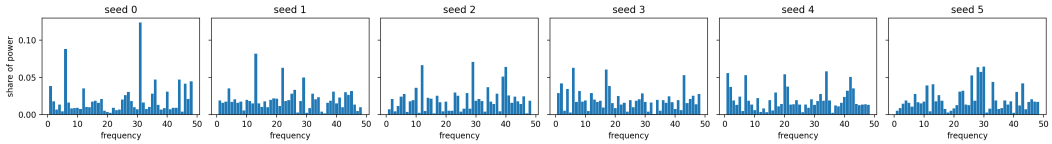
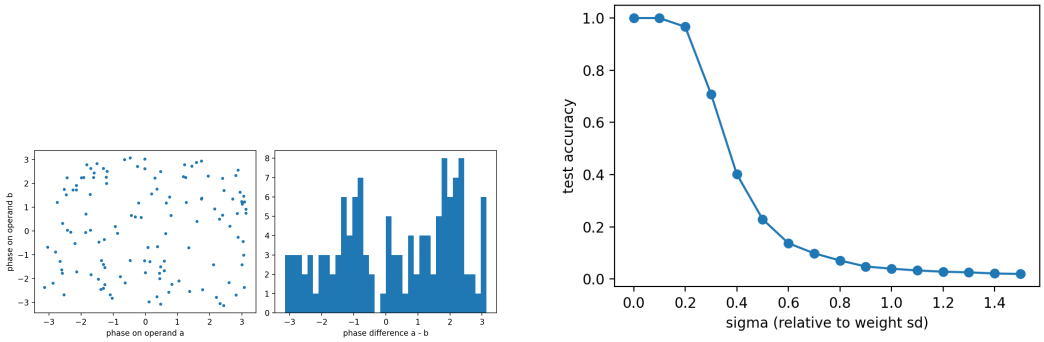


Figure 7: **Fourier power spectra of W_1 for six seeds** (share of power per frequency, f and $p - f$ combined).

Phase structure and robustness. For each of the 128 hidden units, the phases of its dominant frequency on operands a and b are spread over the whole circle, and their differences are close to uniform (Figure 8a). There are no discrete phase families, consistent with the analytical solutions of Gromov [2023], in which each unit's phases are free parameters. Adding Gaussian noise with standard deviation σ times the standard deviation of each weight matrix to W_1 and W_2 lowers test accuracy to 71% at $\sigma = 0.3$, 14% at $\sigma = 0.6$ and 5% at $\sigma = 0.9$ (Figure 8b, seed 0, step 10^5).



(a) Phase alignment.

(b) Weight perturbation robustness.

Figure 8: **Phase structure and robustness of the MLP solution** (seed 0).

E Timing, Replication and Rank Interventions

E.1 Timing of Compression

Table 18 summarizes the timing of compression in both architectures.

Steepest-decrease measure. The steepest decrease used in Section 6 is the step of the most negative central-difference derivative of the effective rank after a centered moving average over five logged steps (`np.convolve` in `same` mode, then `np.gradient`). Each smoothed value belongs to the logged step at its center, so the smoothing adds no delay. The zero-padded average distorts the last two logged steps and the central difference a third, so we exclude the last four. On synthetic signals the measure places a ramp exactly and a sharp step two logging intervals early (Figure 10), so any bias shortens the measured lag. It falls after t_g in 13 of 14 transformer seeds of 0–14 (the exception: 250 steps before t_g) and in 14 of 15 held-out seeds (the fifteenth at t_g). Figure 9 shows the effective rank of W_{in} and the test accuracy of the 14 grokked transformer seeds aligned at t_g. Counting the decay of the initialization, $31 \pm 4\%$ of the decrease of W_{in} happens after t_g (W_E: 22%, W_O: 27%, W_U: 37%, W_{out}: 38%, W_V: 62%).

Table 18: **Timing of compression** (mean $\pm$ s.d., in steps). t_c is the first step at which ER falls below 99% of its running maximum. t_{on} and t_s are the start and the steepest point of the effective-rank drop of W_{in} after the initialization decay. t_{10} is the first step with test accuracy $\geq 10\%$.

	MLP (15 seeds)	Transformer (29 grokked)
Memorization step t_{fit}	$9{,}133 \pm 297$	250 in all seeds
Grokking onset t_g	$12{,}633 \pm 297$	$11{,}914 \pm 4{,}116$
Start of compression $- t_g$	t_c: $6{,}367 \pm 3{,}248$ (15/15 > 0)	t_{on}: $-1{,}888 \pm 992$ (28/29 after t_{10})
Steepest ER decrease $- t_g$	$9{,}667 \pm 3{,}320$ (15/15 > 0)	905 ± 506 (27/29 > 0)
Share of ER decrease after t_g	0.99 ± 0.02	0.63 ± 0.09[a]

[a] From the plateau after the initialization decay. Including the decay, 0.32 ± 0.05.

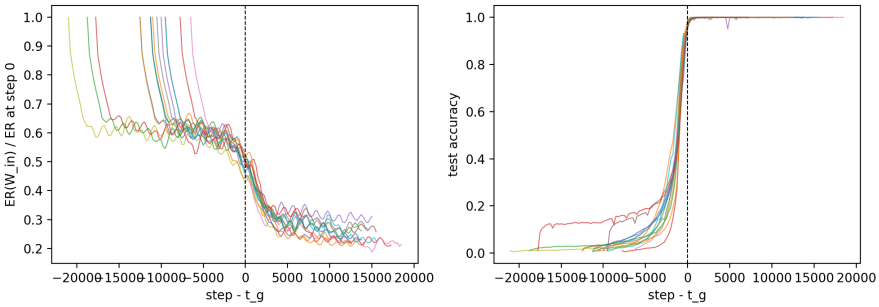


Figure 9: **Effective rank of W_{in}** (relative to step 0) **and test accuracy** of the 14 grokked transformer seeds against steps relative to t_g. The effective rank falls steeply while the random initialization decays, plateaus at about 60% of its initial value, and drops again around t_g.

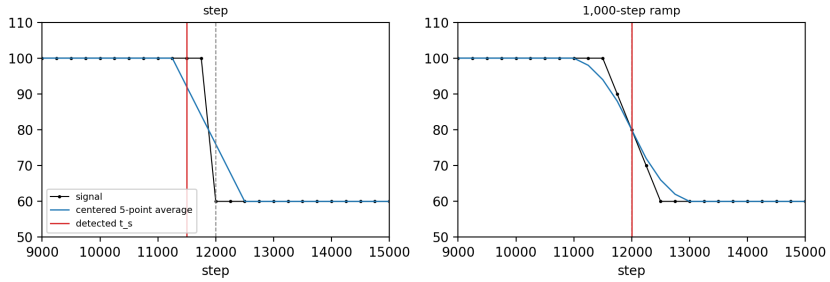
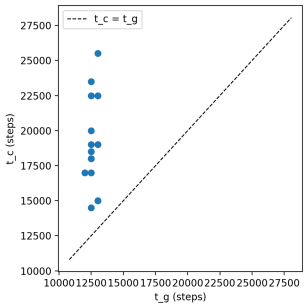
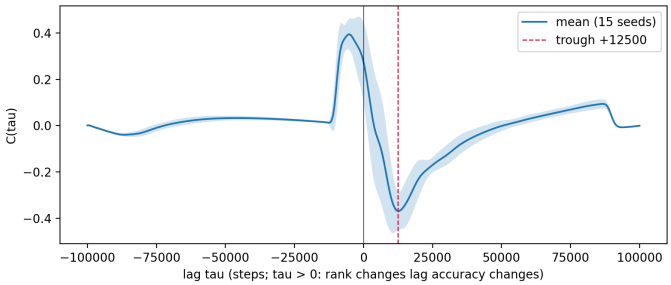


Figure 10: **Synthetic check of the steepest-drop measure** (logging every 250 steps). Left: a sharp step at step 12,000 is located two logging intervals early, because the derivative of the centered average is tied over several steps and the earliest is taken. Right: a 1,000-step ramp centered on step 12,000 is located exactly.

Lead and lag in the MLP. Figure 11 shows that the compression onset follows the grokking onset in every MLP seed, and that the cross-correlation of the derivatives of effective rank and test accuracy has its trough at $+12{,}500$ steps.



(a) t_c vs. t_g: all 15 seeds above the diagonal.



(b) Mean cross-correlation (± 1 s.d.) of $\partial \mathrm{ER}/\partial t$ and $\partial \mathrm{Acc}/\partial t$.

Figure 11: **Compression lags grokking** in every MLP seed.

Granger tests. Granger causality [Granger, 1969] tests whether the past of one time series improves the prediction of another beyond that series' own past. It measures temporal predictive asymmetry, not interventional causation. We use the first differences of effective rank and test accuracy, logged every 500 steps, with lag order $L \leq 4$ chosen by the Akaike information criterion. Augmented Dickey–Fuller tests find these differences stationary in only 8 and 13 of 15 seeds, so we repeat the test after differencing further until the series are stationary in every seed. Changes in test accuracy predict changes in effective rank in every seed in both versions, and the reverse holds in at most 6 of 15 (Table 19). Since test accuracy is flat at 100% for most of training, these tests largely restate the order of the two pulses.

Table 19: **Granger tests** (temporal predictive asymmetry) across 15 MLP seeds.

Series	Null hypothesis	$\bar{F}$	$\bar{p}$	Med. p	Rejected ($p < 0.05$)
First differences	Compression $\not\to$ Grokking	2.84	0.133	0.059	6/15
	Grokking $\not\to$ Compression	24.77	0.003	< 0.001	15/15
Stationary (ADF)	Compression $\not\to$ Grokking	2.44	0.257	0.126	4/15
	Grokking $\not\to$ Compression	19.77	0.003	< 0.001	15/15

E.2 Networks That Memorize Without Grokking

Transformers. Table 20 compares the second effective-rank drop of W_{in} in the 29 grokked transformer seeds with the seven runs that memorize the training set without generalizing. All runs memorize the training set by step 250. In grokked seeds, the effective rank of W_{in} falls by $60 \pm 5\%$ between step 4,000 and the end of training, with a peak rate of 10.4 ± 2.0 units per 1,000 steps. In non-grokked runs it falls by 15–23%, at 2.1–4.6 units per 1,000 steps.

Table 20: **Effective rank of W_{in} after the initialization decay** (transformer). Peak rate: steepest smoothed decrease after step 4,000.

Run	Final test acc. (%)	Peak rate (per 1,000 steps)	Relative change, step 4,000 $\to$ end
Grokked (29 seeds)	≥ 95	-10.4 ± 2.0	-0.60 ± 0.05
wd 1.0, seed 12	8.2	-4.0	-0.15
wd 0.1, seeds 0–2	1.8–4.3	-2.1 to -2.3	-0.22 to -0.23
wd 0.3, seeds 0–2	9.1–13.2	-4.3 to -4.6	-0.17 to -0.19

Start of the drop. The start t_{on} of the drop is the first logged step of the contiguous stretch before t_s in which the smoothed decrease is at least 20% of its peak rate. It falls between the steps at which test accuracy first exceeds 10% and 50% in most seeds: $t_{\text{on}} - t_{10} = 3{,}095 \pm 3{,}778$ steps (28 of 29 non-negative, median 1,750) and $t_{\text{on}} - t_{50} = -828 \pm 964$ (4 of 29 non-negative). Three seeds have

Table 21: **Low rank is acquired at the transition** (transformer, weight decay 1.0). Singular directions carrying 99% of $\sum_i \sigma_i^2$ at the same training step, for grokked seeds before and after their own onset and for the seed that never groks. Mean $\pm$ s.d., with the number of seeds in parentheses.

Step	Group	W_E	W_{in}	W_{out}	W_U
10,000	before own onset (17)	47.8 ± 2.4	77.2 ± 8.0	64.1 ± 4.5	60.4 ± 4.4
	after own onset (12)	39.0 ± 4.6	42.8 ± 12.3	33.2 ± 12.3	34.8 ± 7.1
	never groks (1)	52	84	71	67
12,500	before own onset (9)	44.3 ± 2.8	72.2 ± 7.4	61.7 ± 3.6	60.2 ± 3.7
	after own onset (20)	36.0 ± 4.1	32.1 ± 13.3	24.2 ± 15.1	30.4 ± 9.5
	never groks (1)	50	86	70	66
15,000	before own onset (8)	43.5 ± 2.6	67.2 ± 8.5	59.1 ± 3.2	57.2 ± 2.7
	after own onset (21)	32.1 ± 4.8	22.4 ± 3.4	15.0 ± 5.1	24.7 ± 5.3
	never groks (1)	45	81	67	65

$t_{10} = 500$ from an early fluctuation of test accuracy, which inflates the mean of the first difference. Test accuracy at t_{on} is $34 \pm 22\%$, and 37% of the post-plateau decrease happens before t_g.

Matched steps. Table 21 gives the variance counts at steps 10,000–15,000. At step 5,000, before any seed has generalized, the 29 grokked seeds and the non-grokked seed at the same weight decay have nearly the same counts (73.5 ± 3.9 against 75 for W_{in}, 65.2 ± 2.1 against 70 for W_U), while the runs at weight decay 0.1 and 0.3 have higher counts (113 and 92 for W_{in}), a weight-decay effect. At step 20,000, 28 seeds have generalized and their counts are 32.3 ± 4.7, 21.4 ± 7.0, 15.2 ± 7.0 and 24.5 ± 5.3 for W_E, W_{in}, W_{out} and W_U, against 48, 78, 68 and 64 for the non-grokked seed.

MLPs trained on random labels. Figure 12 shows the effective rank and training accuracy of the five random-label MLPs next to the grokked ones. The random-label networks reach 90% training accuracy at step $7{,}300 \pm 274$, end at $84 \pm 2\%$, and their effective rank falls from 115.0 ± 0.1 to 113.7 ± 0.4 ($1.1 \pm 0.3\%$), against 116.1 ± 0.4 to 104.0 ± 2.1 ($10.5 \pm 1.8\%$) in grokked networks.

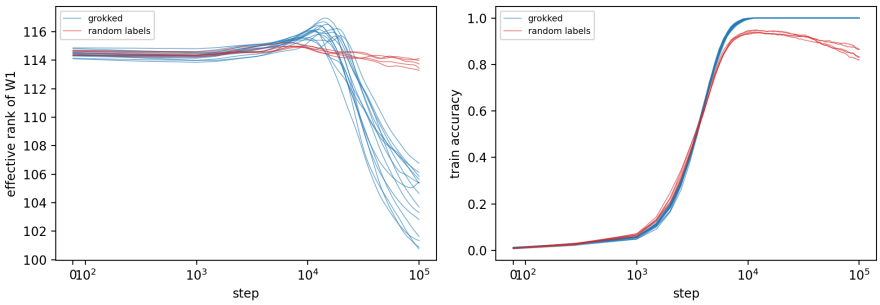


Figure 12: **MLPs that memorize without generalizing barely compress.** Effective rank of W_1 (left) and training accuracy (right) for 15 grokked MLPs (blue) and 5 random-label MLPs (red).

E.3 Replication on Held-Out Seeds and Tasks

Data. H1 uses the transformer exactly as in the main experiments (weight decay 1.0, 25,000 steps) with seeds 15–29, which were trained only after the timing measure had been defined on seeds 0–14. H2–H5 use the MLP exactly as in the main runs, on $ab \bmod 97$ and $(a{+}b) \bmod 113$, with seeds 0–14 per task. H6 uses the same MLP with $W_1 = AB$, $r \in \{128, 64, 32\}$, seeds 0–2 per task and rank. The analysis scripts are in the code repository.

Definitions. t_g is the first logged step with test accuracy $\geq 95\%$. A run counts as grokked if its final test accuracy is $\geq 95\%$, and only grokked runs enter the tests. t_s is the logged step with the most negative derivative of the five-point moving average of the effective rank, over logged steps from step 4,000 (transformer, W_{in}) or 1,000 (MLP, W_1) up to four logging intervals before the end. t_c is the first logged step $\geq 1{,}000$ at which the effective rank of W_1 falls below 99% of its running

maximum. $d_{0.99}(s)$ is the smallest k whose top-k truncation of W_1 keeps $\geq 99\%$ of the test accuracy of checkpoint s. s_g is the first saved checkpoint $\geq t_g$, s_+ the first saved checkpoint $\geq t_g + 5{,}000$ that is not the final one, and s_{end} the final checkpoint.

Tests. All tests are one-sided at $\alpha = 0.05$ and run separately per task, and every result is reported. H1: $t_s - t_g > 0$ (one-sample t-test, with the number of seeds with $t_s > t_g$ and an exact sign test). H2: $t_c - t_g > 0$ and H3: $t_s - t_g > 0$ (one-sample t-tests). H4: $d_{0.99}(s_g) > d_{0.99}(s_+)$ and H5: $d_{0.99}(s_{\text{end}}) > d_{0.99}(s_+)$ (paired t-tests). H6 (descriptive): the factorized full-rank model groks earlier than the unfactorized one, and t_g increases as r decreases from 128 to 64 to 32.

Results. Every test is significant at $\alpha = 0.05$ (Tables 22, 24 and 23, and Table 27 for H6). With a Bonferroni correction over the nine t-tests, H4 on $(a+b) \bmod 113$ is not. For H1, the 95% confidence interval of $t_s - t_g$ is $[560, 1{,}140]$ steps (median 750), and the exact sign test gives $p = 4.9 \times 10^{-4}$. Twice the number of frequencies holding 99% of the Fourier power of W_1 is 90.0 ± 2.4 for $ab \bmod 97$ (Fourier transform over the nonzero residues ordered by the discrete logarithm base 5) against $d_{0.99} = 89.8 \pm 2.0$, but 105.5 ± 2.1 for $(a+b) \bmod 113$ against $d_{0.99} = 85.5 \pm 3.2$, so the frequency count does not predict $d_{0.99}$ in general (§8). Multiplication loses a little test accuracy late in training, from 100.0% at step 30,000 to $99.7 \pm 0.3\%$ at the end. d_{out}, which H4 and H5 do not use, does not narrow on either task: it is 121.7–127.4 at every checkpoint (Table 16).

Table 22: **Replication tests** (one-sided tests, mean $\pm$ s.d. in steps or directions, with the number of seeds showing a positive difference in parentheses).

	Quantity	Transformer, held-out	MLP, $ab \bmod 97$	MLP, $(a+b) \bmod 113$
H1	$t_s - t_g$ (W_{in})	850 ± 524 (14/15)*	–	–
H2	$t_c - t_g$	–	$6{,}300 \pm 1{,}461$ (15/15)*	$6{,}700 \pm 4{,}519$ (15/15)*
H3	$t_s - t_g$ (W_1)	–	$9{,}800 \pm 2{,}274$ (15/15)*	$11{,}867 \pm 6{,}741$ (15/15)*
H4	$d_{0.99}(s_g) - d_{0.99}(s_+)$	–	32.9 ± 13.7 (15/15)*	10.8 ± 16.8 (12/15)*
H5	$d_{0.99}(s_{\text{end}}) - d_{0.99}(s_+)$	–	25.1 ± 2.9 (15/15)*	16.9 ± 3.1 (15/15)*
H6	t_g for AB at $r = 128/64/32$, and unfactorized	–	2,000/2,667/4,667 and 12,833	2,000/3,000/4,500 and 15,333

*$p < 0.05$. The largest p is 0.013 (H4, $(a+b) \bmod 113$). H1–H3 use one-sample t-tests, H4 and H5 paired t-tests, and H6 is descriptive, with 3 seeds per arm. s_g is the first checkpoint at or after t_g, s_+ the first checkpoint at least 5,000 steps later, and s_{end} the final checkpoint.

Table 23: **H2–H5 on the two further MLP tasks** (15 seeds each, mean $\pm$ s.d.). Below each test, the number of seeds with a positive difference, the t statistic and the one-sided p.

	$ab \bmod 97$	$(a+b) \bmod 113$
Grokked	15/15	15/15
t_g	$12{,}633 \pm 297$	$15{,}567 \pm 417$
t_c	$18{,}933 \pm 1{,}498$	$22{,}267 \pm 4{,}629$
H2: $t_c - t_g$	$6{,}300 \pm 1{,}461$ 15/15, $t = 16.70$, $p = 6.1 \times 10^{-11}$	$6{,}700 \pm 4{,}519$ 15/15, $t = 5.74$, $p = 2.6 \times 10^{-5}$
H3: $t_s - t_g$	$9{,}800 \pm 2{,}274$ 15/15, $t = 16.69$, $p = 6.2 \times 10^{-11}$	$11{,}867 \pm 6{,}741$ 15/15, $t = 6.82$, $p = 4.2 \times 10^{-6}$
H4: $d_{0.99}(s_g) - d_{0.99}(s_+)$	32.9 ± 13.7 15/15, $t = 9.31$, $p = 1.1 \times 10^{-7}$	10.8 ± 16.8 12/15, $t = 2.49$, $p = 0.013$
H5: $d_{0.99}(s_{\text{end}}) - d_{0.99}(s_+)$	25.1 ± 2.9 15/15, $t = 33.01$, $p = 5.6 \times 10^{-15}$	16.9 ± 3.1 15/15, $t = 20.98$, $p = 2.8 \times 10^{-12}$
Share of the ER(W_1) decrease after t_g	1.00 ± 0.01	0.98 ± 0.02
$d_{0.99}$ at the end	89.8 ± 2.0	85.5 ± 3.2
Top / bottom / random-subspace truncation (%)	99.10 / 36.84 / 20.68	99.08 / 27.93 / 33.03

Truncations keep $k = d_{0.99}$ directions. Chance accuracy: 1.03% ($p = 97$) and 0.88% ($p = 113$). For $(a+b) \bmod 113$, s_g is step 20,000 in 12 seeds. The three seeds with s_g at step 15,000 narrow from 106, 111 and 115 directions to 66, 71 and 67.

Table 24: **H1 per seed:** held-out transformer seeds, t_g and $t_s - t_g$ (steps).

Seed	t_g	$t_s - t_g$	Seed	t_g	$t_s - t_g$	Seed	t_g	$t_s - t_g$
15	7,000	+500	20	17,500	+500	25	17,750	+250
16	12,500	+1,000	21	8,000	0	26	8,750	+500
17	7,500	+750	22	13,000	+1,000	27	15,750	+750
18	17,000	+1,000	23	16,500	+1,750	28	8,500	+750
19	7,500	+750	24	10,250	+2,000	29	8,250	+1,250

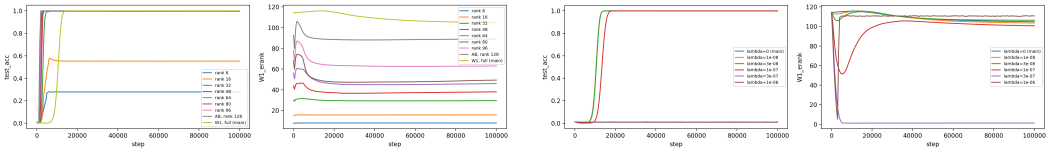
E.4 Forcing Low Rank

Table 25 and Figure 13 report the MLP rank interventions summarized in Section 6.4. Table 26 shows that the factorization alone biases W_1 toward low rank, and Table 27 repeats the factorization on the two further tasks. Table 28 and Figure 14 repeat the factorized control and two rank caps with logging every 50 steps, on the same seeds as above. The full-rank factorization groks 7.5 ± 0.1, 7.2 ± 0.1 and 8.4 ± 0.2 times sooner than the unfactorized model on the three tasks. Its effective rank dips early, rises to its maximum at the onset and falls afterwards.

Table 25: **Forcing low rank from initialization** (MLP). Rank-constrained models use 3 seeds and the penalty 5 seeds.

Model	Grokked	t_g	Final test acc. (%)	Final ER(W_1)
Unfactorized W_1 (main)	3/3	$12{,}833 \pm 289$	100.0 ± 0.0	104.9 ± 1.1
$W_1 = AB$, $r = 128$ (control)	3/3	$2{,}000 \pm 0$	99.9 ± 0.2	89.1 ± 0.9
$W_1 = AB$, $r = 96$	3/3	$2{,}167 \pm 289$	100.0 ± 0.0	63.0 ± 5.7
$W_1 = AB$, $r = 80$	3/3	$2{,}500 \pm 0$	100.0 ± 0.0	49.4 ± 1.2
$W_1 = AB$, $r = 64$	3/3	$2{,}667 \pm 289$	100.0 ± 0.0	46.0 ± 3.7
$W_1 = AB$, $r = 48$	3/3	$3{,}167 \pm 289$	100.0 ± 0.0	38.0 ± 0.6
$W_1 = AB$, $r = 32$	3/3	$4{,}500 \pm 0$	99.5 ± 0.3	29.5 ± 0.1
$W_1 = AB$, $r = 16$	0/3	–	55.3 ± 10.6	15.8 ± 0.0
$W_1 = AB$, $r = 8$	0/3	–	27.8 ± 6.1	8.0 ± 0.0
No penalty	5/5	$12{,}700 \pm 447$	100.0 ± 0.0	104.9 ± 1.6
$\lambda = 10^{-8}$	5/5	$12{,}800 \pm 274$	100.0 ± 0.0	103.5 ± 2.2
$\lambda = 3 \times 10^{-8}$	5/5	$12{,}800 \pm 274$	100.0 ± 0.1	106.3 ± 1.3
$\lambda = 10^{-7}$	5/5	$16{,}200 \pm 447$	100.0 ± 0.0	100.7 ± 3.4
$\lambda = 3 \times 10^{-7}$	0/5	–	0.8 ± 0.1	1.2 ± 0.3
$\lambda = 10^{-6}$	0/5	–	1.1 ± 0.1	110.8 ± 0.4[§]

[§]At $\lambda = 10^{-6}$ the penalty prevents learning. The high effective rank presumably reflects weights shrunk to small, unstructured values, which we did not inspect.



(a) Rank-constrained parameterization $W_1 = AB$.

(b) Nuclear-norm penalty from step 0.

Figure 13: **Forcing low rank from initialization** (mean test accuracy and effective rank of W_1 against training step).

Table 26: **The factorization biases W_1 toward low rank.** Effective rank of $W_1 = AB$ with $r = 128$ and of the unfactorized W_1 (seeds 0–2, $(a+b) \bmod 97$). The factorized model groks at step 2,000, the unfactorized one at $12{,}833 \pm 289$.

Step	$W_1 = AB$, $r = 128$	Unfactorized W_1
0	92.7 ± 0.3	114.6 ± 0.1
500	79.7 ± 1.0	114.5 ± 0.1
1,000	97.4 ± 0.6	114.5 ± 0.1
1,500	105.5 ± 0.3	114.6 ± 0.2
2,000	105.9 ± 0.3	114.7 ± 0.2
3,000	102.4 ± 2.0	114.9 ± 0.2
5,000	94.8 ± 3.3	115.1 ± 0.1
10,000	89.8 ± 2.7	115.9 ± 0.1
20,000	88.5 ± 2.1	115.1 ± 0.4

Table 27: **Factorization runs on the two further tasks (H6)** (3 seeds per arm, unfactorized: main seeds 0–2).

Task	Model	Grokked	t_g	Final test acc. (%)	Final ER(W_1)
$ab \bmod 97$	Unfactorized W_1	3/3	$12{,}833 \pm 289$	99.7 ± 0.4	102.3 ± 0.7
	$W_1 = AB$, $r = 128$	3/3	$2{,}000 \pm 0$	99.3 ± 0.7	85.6 ± 1.4
	$W_1 = AB$, $r = 64$	3/3	$2{,}667 \pm 289$	100.0 ± 0.0	47.9 ± 0.2
	$W_1 = AB$, $r = 32$	3/3	$4{,}667 \pm 289$	99.0 ± 0.5	30.1 ± 0.2
$(a+b) \bmod 113$	Unfactorized W_1	3/3	$15{,}333 \pm 289$	100.0 ± 0.0	109.5 ± 1.8
	$W_1 = AB$, $r = 128$	3/3	$2{,}000 \pm 0$	100.0 ± 0.0	89.9 ± 2.1
	$W_1 = AB$, $r = 64$	3/3	$3{,}000 \pm 0$	100.0 ± 0.0	47.0 ± 5.1
	$W_1 = AB$, $r = 32$	3/3	$4{,}500 \pm 0$	99.2 ± 0.9	29.6 ± 0.5

Table 28: **Factorized MLPs logged every 50 steps** (3 seeds each, mean $\pm$ s.d.). ER: effective rank of $W_1 = AB$.

Task	Model	t_g	ER minimum before t_g (step)	ER at t_g	ER at step 8,000
$(a+b) \bmod 97$	unfactorized	$12{,}483 \pm 340$	–	116.2 ± 0.2	–
	$r = 128$	$1{,}667 \pm 58$	76.9 ± 0.7 (333)	106.3 ± 0.5	90.7 ± 3.1
	$r = 64$	$2{,}450 \pm 100$	49.3 ± 0.2 (267)	60.5 ± 0.1	52.5 ± 6.1
	$r = 32$	$4{,}133 \pm 29$	28.2 ± 0.2 (300)	31.5 ± 0.1	31.0 ± 0.2
$ab \bmod 97$	unfactorized	$12{,}567 \pm 189$	–	115.6 ± 0.6	–
	$r = 128$	$1{,}750 \pm 0$	73.8 ± 0.6 (300)	106.0 ± 1.3	87.3 ± 1.5
$(a+b) \bmod 113$	unfactorized	$15{,}233 \pm 293$	–	117.6 ± 0.2	–
	$r = 128$	$1{,}817 \pm 29$	78.6 ± 0.0 (333)	108.5 ± 0.5	93.4 ± 2.3

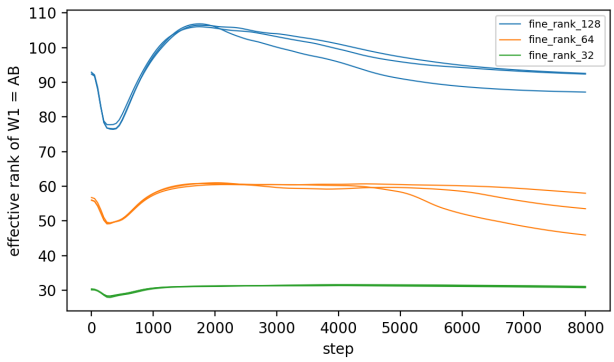


Figure 14: **Effective rank of factorized MLPs** ($W_1 = AB$ with $r = 128$, 64 and 32, logged every 50 steps, 3 seeds each).

E.5 Label-Free Onset Detection

Table 29 reports the onset detectors summarized in Section 6.5, and Figure 16 the Fourier IPR of one MLP seed. In the MLP, the Fourier IPR fires at or just after the end of the baseline window, so its lead is set by the window, $t_g - W$. Spectral IPR, effective rank and stable rank of W_1 change after t_g (early in at most 2 of 15 seeds, with median leads from $-8{,}000$ to $-23{,}500$ steps). For the transformer, "early" is counted out of the seeds with $t_g > W$, the only ones that can be early. Of 63 metric, window and threshold combinations tested on the MLP, only the gradient norm (window 5,000, threshold 3σ) detected grokking early in all seeds without false alarms. Given the number of combinations, we regard this as exploratory.

Table 29: **Label-free onset detection** (threshold 3σ above the mean of the baseline window $[0, W)$). *At window end*: fires at the first logged step after the window. *False alarms*: fires in runs that never generalize (5 random-label MLP runs and 7 non-grokked transformer runs). Lead $= t_g - t_{\text{onset}}$ (median, steps).

Model	Metric	W	Early	At window end	False alarms	Lead
MLP	Fourier IPR (W_1)	2,500	15/15	14/15	5/5	10,000
		5,000	15/15	15/15	5/5	7,500
		10,000	15/15	14/15	5/5	2,500
MLP	Gradient norm	2,500	15/15	0/15	5/5	9,000
		5,000	15/15	0/15	0/5	2,500
		10,000	3/15	0/15	3/5	0
Transformer	Fourier IPR (W_E)	2,500	14/14	0/14	7/7	7,625
		5,000	14/14	0/14	7/7	5,250
		10,000	9/9	6/14	7/7	1,000
Transformer	Spectral IPR (W_{in})	2,500	14/14	0/14	7/7	3,125
		5,000	14/14	0/14	7/7	1,625
		10,000	9/9	5/14	3/7	750

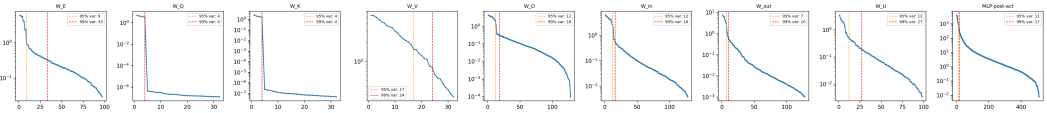


Figure 15: **Transformer singular value spectra** (seed 0) for W_E, W_Q, W_K, W_V, W_O, W_{in}, W_{out}, W_U and the MLP post-activations, with the 95% and 99% variance thresholds.

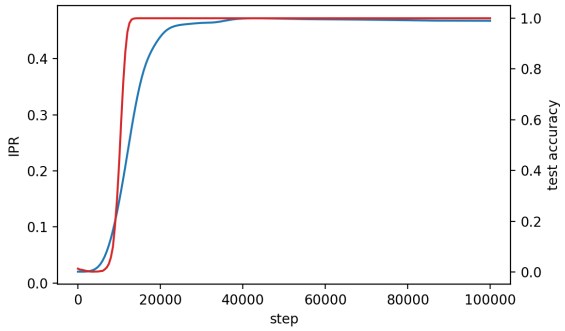


Figure 16: **Fourier IPR of** W_1 (blue) and test accuracy (red), MLP seed 0.

F Transformer: Additional Results

F.1 Dimensionality at the End of Training

Table 30 gives the variance counts and $d_{0.99}$ of every analyzed matrix, and Figure 15 the singular value spectra for one seed. Averaged over seeds at fixed steps, $d_{0.99}$ mixes seeds before and after their onsets: $d_{0.99}(W_{\text{in}})$ falls from 18.2 ± 7.2 at step 10,000 (mean test accuracy $57 \pm 40\%$) to 12.4 ± 1.5 at step 20,000, and $d_{0.99}(W_E)$ from 19.4 ± 9.9 to 7.7 ± 1.5. Aligned at each seed's own onset (Table 5), d_{out} narrows during the rise in test accuracy. Table 31 compares the final counts with those of the MLP.

Table 30: **Transformer dimensionality** (14 grokked seeds, final checkpoint). k_{95}, k_{99}: singular directions carrying 95% and 99% of $\sum_i \sigma_i^2$. $d_{0.99}$: fewest directions that keep 99% of test accuracy. Attention projections are stacked across the four heads. Activations are truncated after mean-centering.

Matrix	Max rank	k_{95}	k_{99}	$d_{0.99}$
W_E (embedding)	97	8.8 ± 1.5	30.4 ± 4.0	7.7 ± 1.5
W_Q / W_K	32	4.0 ± 0.0	4.0 ± 0.0	2.6 ± 0.5 / 2.4 ± 0.7
W_V	32	18.5 ± 2.6	24.3 ± 2.6	10.9 ± 1.5
W_O	128	13.2 ± 1.4	21.3 ± 4.7	12.8 ± 1.3
W_{in}	128	13.1 ± 1.4	18.1 ± 3.8	13.0 ± 1.3
W_{out}	128	8.6 ± 1.4	11.8 ± 2.5	8.7 ± 1.5
W_U (unembedding)	98	10.3 ± 1.3	23.8 ± 2.8	8.8 ± 1.4
MLP post-activations	512	14.4 ± 3.3	23.6 ± 7.2	17.2 ± 3.3

Table 31: **Architecture comparison** (final checkpoints, with 15 MLP seeds, 5 cross-entropy MLP seeds per learning rate and 14 grokked transformer seeds).

Property	MLP (MSE)	MLP (CE)	Transformer
Loss, weight decay, training steps	MSE, 10^{-2}, 10^5	CE, 10^{-2}, 10^5	CE, 1.0, 2.5×10^4
$d_{0.99}$ of first-layer weights (of 128)	89.1 ± 2.3	48–51	13.0 ± 1.3
d_{out} of first-layer weights (of 128)	127.3 ± 0.6	70–81	13.0 ± 1.9
When d_{out} falls	never	late, after t_g	during the transition
Frequencies holding 90% of first-layer power	33.7 ± 1.9 (W_1)	–	4.2 ± 0.7 (W_E)

F.2 Dense Time Course

All transformer runs were repeated with a checkpoint every 250 steps. For each grokked seed we compute d_{out} of W_E, W_{in}, W_{out} and W_U at every checkpoint within 3,000 steps of its t_g, testing every k from 1 upward. The restricted and excluded losses follow Nanda et al. [2023] in output space: for each input, we keep (restricted) or remove (excluded) the components of the centered residue logits along $\cos(2\pi wc/p)$ and $\sin(2\pi wc/p)$ for the five key frequencies w of W_E, and evaluate the

cross-entropy on the training set. For each curve, the midpoint is the first checkpoint at which the count falls below the midpoint between its value at the start of the window and its final value. The restricted loss uses the same rule, and the excluded loss the first checkpoint above the midpoint of its rise. The midpoints relative to t_g are: W_E $+26 \pm 488$ steps (11 of 29 seeds before t_g), W_{in} -862 ± 503 (29 of 29), W_{out} -957 ± 328 (29 of 29), W_U -741 ± 375 (28 of 29), restricted loss $-1{,}966 \pm 647$ and excluded loss $-1{,}026 \pm 646$.

Table 32 gives d_{out} for the runs that never generalize. They narrow slowly, most at weight decay 0.3, where test accuracy also creeps up to 9–13%, but at the end of training they remain two to nine times wider than the grokked seeds.

Table 32: **Runs that never generalize**: d_{out} (centered logits within 10%, test inputs) at step 5,000 and at the end of training, with grokked seeds 3,000 steps after their onset for comparison.

Run	Step	W_E	W_{in}	W_{out}	W_U	Test acc. (%)
wd 1.0, seed 12	5,000	31	33	64	70	2.8
	end	23	26	49	55	8.2
wd 0.1, seeds 0–2	5,000	45–47	42–43	74–77	81–84	1.0–2.8
	end	35–38	35–37	67–69	78–83	1.8–4.3
wd 0.3, seeds 0–2	5,000	40–42	38–40	71–73	79–82	1.2–2.9
	end	29–30	30–31	48–52	53–59	9.1–13.2
Grokked (29), $t_g + 3{,}000$	–	8.6 ± 2.0	13.8 ± 1.7	9.0 ± 1.8	8.9 ± 1.7	100.0

F.3 Runs That Never Grok and Attention Projections

Table 33 lists the seven transformer runs that did not reach 95% test accuracy within 25,000 steps. W_Q has rank 4 in all of them, and W_K in those trained with weight decay of at least 0.3. At weight decay 0.1, where the random initialization has not yet decayed, W_K keeps 18–19 directions. The rank-4 structure of W_Q and W_K, one direction per head, is architectural rather than a signature of grokking: only the attention pattern of the final position, which always holds the "=" token, affects the output, so each head's query is a single vector and the gradients of its query and key projections have rank one.

Table 33: Non-grokked transformer runs: final accuracy (%) and k_{99} (singular directions carrying 99% of $\sum_i \sigma_i^2$). Grokked seeds (mean): W_E 30.4, W_Q 4.0, W_K 4.0, W_V 24.3, W_O 21.3, W_{in} 18.1, W_{out} 11.8, W_U 23.8.

Run	Train	Test	W_E	W_Q	W_K	W_V	W_O	W_{in}	W_{out}	W_U
wd 1.0, seed 12	100.0	8.2	42	4	4	30	42	68	62	62
wd 0.1, seed 0	100.0	4.3	43	4	19	31	50	82	78	72
wd 0.1, seed 1	100.0	1.8	43	4	18	31	51	81	78	73
wd 0.1, seed 2	100.0	3.0	44	4	18	31	52	84	79	73
wd 0.3, seed 0	100.0	13.2	34	4	4	31	43	74	61	62
wd 0.3, seed 1	100.0	10.7	36	4	4	31	43	73	61	62
wd 0.3, seed 2	100.0	9.1	37	4	4	30	43	77	65	65

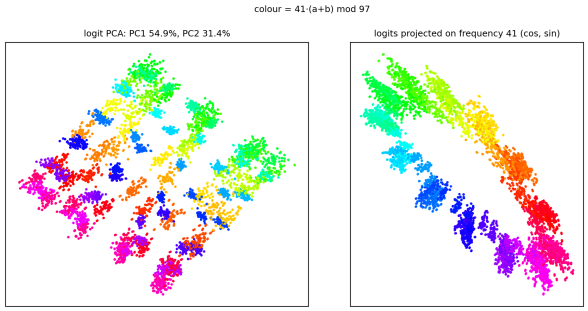
F.4 Residual-Stream Alignment

Consecutive components communicate through shared subspaces of the 128-dimensional residual stream, but this is not specific to grokking (Table 34). What attention writes is almost exactly what the MLP reads (0.923 ± 0.035), and what the MLP writes is what the unembedding reads (0.885 ± 0.078), whereas the MLP's own read and write subspaces are only weakly aligned (0.330 ± 0.050). The attention-to-MLP alignment is equally high in the non-grokked runs (0.839–0.947). For a matrix that reads from the residual stream (W_{in}, W_U), its read subspace is spanned by its top-k left singular vectors. For a matrix that writes to it (W_O, W_{out}), its write subspace is spanned by its top-k right singular vectors. For two subspaces with orthonormal bases $U, V \in \mathbb{R}^{128 \times k}$, the principal angles are the arccosines of the singular values of $U^\top V$, and we report the mean of their cosines, with $k = 11 = 2 \times 5$ key frequencies $+1$. The key-frequency subspace of W_E is spanned by the images

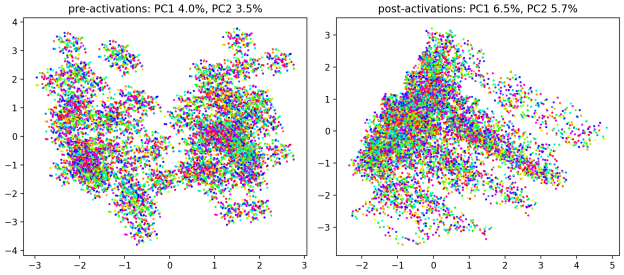
under $W_E^\top$ of the cosine and sine of each of its five strongest frequencies and of the constant vector. Random baselines average 300 pairs of independent random subspaces.

Table 34: **Residual-stream alignment** (mean cosine of principal angles between 11-dimensional subspaces). Not grokked: three seeds each at weight decay 0.1 and 0.3, and the one non-grokked seed at weight decay 1.0.

Subspace pair	Grokked	Not grokked		
	wd 1.0 ($n = 14$)	wd 0.1 ($n = 3$)	wd 0.3 ($n = 3$)	wd 1.0 ($n = 1$)
Attention write (W_O) vs. MLP read (W_{in})	0.923 ± 0.035	0.839 ± 0.021	0.918 ± 0.048	0.947
MLP read (W_{in}) vs. MLP write (W_{out})	0.330 ± 0.050	0.269 ± 0.024	0.271 ± 0.043	0.218
MLP write (W_{out}) vs. unembedding read (W_U)	0.885 ± 0.078	–	–	–
MLP read vs. key frequencies of W_E	0.252 ± 0.018	–	–	–
MLP write vs. key frequencies of W_E	0.254 ± 0.019	–	–	–
Random baseline	0.248 ± 0.017			



(a) Transformer logits: first two principal components (left) and projection on the dominant logit frequency (right), colored by $41(a + b) \bmod 97$.



(b) MLP pre-activations (left) and post-activations (right): first two principal components, colored by $(a + b) \bmod 97$.

Figure 17: **Representational geometry** (seed 0).

F.5 Representational Geometry

Transformer logits are dominated by a few frequencies, while the MLP's representations are not (Figure 17). Five frequencies carry 87.9% of the logit power (seed 0, after removing each input's mean logit). The first two principal components of the transformer's logits carry 54.9% and 31.4% of their variance and show clusters of nearby classes, and projected on the strongest frequency ($f^\star = 41$) the logits lie on a ring ordered by $41(a + b) \bmod 97$. The first two principal components of the MLP's pre- and post-activations carry only 4.0%/3.5% and 6.5%/5.7% of the variance and show no ring, consistent with the dozens of frequencies the MLP uses (Section 8).