Detecting overfitting in Neural Networks during long-horizon grokking using Random Matrix Theory

author names withheld

Under Review for the Workshop on High-dimensional Learning Dynamics, 2026

Abstract

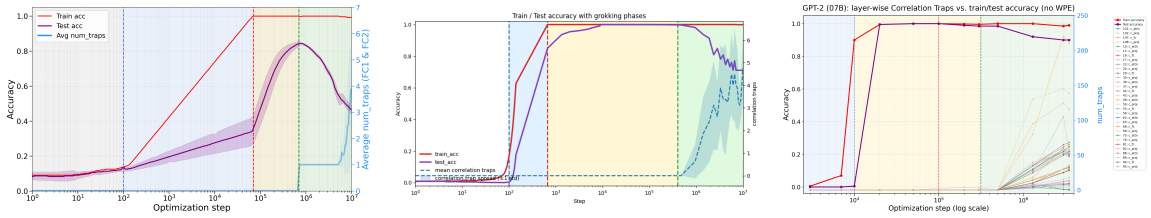
Training Neural Networks (NNs) without overfitting is difficult; detecting that overfitting is difficult as well. We present a novel Random Matrix Theory method that detects the onset of overfitting in deep learning models without access to train or test data. For each model layer, we randomize each weight matrix element-wise, $\mathbf{W} \to \mathbf{W}_{\mathrm{rand}}$, fit the shuffled matrix's empirical spectral distribution with a Marchenko-Pastur distribution, and identify large outliers that violate self-averaging. We call these outliers Correlation Traps. During the onset of overfitting, which we call the "anti-grokking" phase in long-horizon grokking, Correlation Traps form and grow in number and scale as test accuracy decreases while train accuracy remains high. Traps may be benign or may harm generalization; we provide an empirical approach to distinguish between them by passing random data through the trained model and evaluating the JS divergence of output logits. Our findings show that anti-grokking is an additional grokking phase with high train accuracy and decreasing test accuracy, structurally distinct from pre-grokking through its Correlation Traps. More broadly, we find that some foundation-scale LLMs exhibit the same Correlation Traps, indicating potentially harmful overfitting.

1. Introduction

Open-weight models are increasingly used as foundations for downstream systems, but it is often unclear whether a checkpoint is robustly trained or overfit to its training distribution. Users may have access to the weights and a model card, but not to the training data, held-out losses, optimizer state, or long-horizon checkpoint history. As a result, two models can look similar from the outside while differing sharply in whether their weights encode useful structure or brittle, data-specific correlations. This motivates a basic question: can we detect signatures of overfitting directly from the weights of a trained model, without having access to the training or any test data?

Because such histories are usually unavailable for open-weight checkpoints, we first study a setting where the relevant dynamics are visible and overfitting can be readily induced with long-horizon training: *grokking*. In grokking, training accuracy reaches near perfection while test accuracy stays near chance for many optimization steps, before abruptly improving [19]. Grokking has been studied across several architectures and tasks, including algorithmic tasks such as modular addition, computer vision models, and GPT-style transformers [16, 21]. We extend this view to the long-horizon after grokking, where continued training can drive the model into a classical overfitting phase. We call this post-generalization regime *anti-grokking*.

Our setup lets us compare three phases of learning: pre-grokking, grokking, and anti-grokking. Pre-grokking and anti-grokking can look deceptively similar from train and test accuracy alone: in



(a) MLP (b) Modular Addition (MA) (c) GPT2

Figure 1: **Grokking dynamics and Correlation Traps.** Train accuracy, test accuracy, and Correlation Traps are shown for (a) an MLP on MNIST, (b) Modular Addition, and (c) GPT2. Shaded regions denote pre-grokking, grokking, and anti-grokking. Correlation Traps arise at the onset of late-stage test-accuracy decline and track it as it decreases.

both regimes, training accuracy is high while held-out accuracy is poor. But they are structurally different. Pre-grokking occurs before the model has found a generalizing solution; anti-grokking occurs after such a solution has been found and then lost. This makes long-horizon grokking an excellent case for studying harmful overfitting.

To detect this post-grokking overfitting structure, we introduce *Correlation Traps*. Correlation Traps are outliers to the self-averaging Random Matrix Theory (RMT) description of layer weight matrices. For each layer, we shuffle the weights entry-wise, $\mathbf{W} \to \mathbf{W}^{\mathrm{rand}}$, fit the eigenvalue spectrum of $\mathbf{X}^{\mathrm{rand}} = N^{-1}(\mathbf{W}^{\mathrm{rand}})^{\top}\mathbf{W}^{\mathrm{rand}}$ to a Marchenko–Pastur (MP) bulk, and count outliers far beyond the MP edge. The key observation is that while both pre-grokking and anti-grokking can look similar, only anti-grokking exhibits Correlation Traps.

Figure 1 shows the phenomenon in three controlled settings. In all three cases, training separates into three phases. During pre-grokking, training accuracy increases while test accuracy remains low. During grokking, test accuracy improves and the model reaches a high-generalization solution. During anti-grokking, test accuracy falls despite sustained near-perfect training accuracy. Across all three tasks, the trap signal tracks this late divergence: trap counts are low while generalization improves and increase as test accuracy declines. These results indicate that Correlation Traps reveal the onset of an overfitting phase that train/test curves alone would not distinguish from pre-grokking.

Our Contributions. We define Correlation Traps as outliers to the MP/TW right edge in shuffled layer spectra; show that trap onset tracks anti-grokking in MLP, MA, and GPT2-style runs; link traps to non-self-averaging through localization and condensation; distinguish behaviorally active from benign traps using a data-free JSD ablation test; and screen GPT-OSS 20B/120B checkpoints with layer-wise trap profiles. Most importantly, we argue that Correlation Traps can be used to detect signatures of potentially harmful overfitting in open-source, foundation-scale models.

2. Related Work

Grokking and long-horizon generalization. Grokking was introduced by Power et al. [19] as delayed generalization after training accuracy has already saturated. Subsequent work has studied grokking in algorithmic and mechanistic settings, including modular arithmetic, MLPs, and Transformer models [9, 10, 16, 21]. Most of this literature focuses on the transition from memorization to generalization: the model first fits the training set, then later discovers a rule that

generalizes. Our focus is the long-horizon regime after this transition, where a model that has already grokked can lose held-out performance and overfit its training data after extended optimization.

Random-matrix diagnostics of neural-network weights. Our diagnostic builds on spectral approaches to neural-network weight matrices, especially RMT and the `WeightWatcher` framework [12–15]. Prior spectral diagnostics use heavy-tailed structure of the correlated, unrandomized layer weight matrices $\mathbf{W}$, and related metrics to characterize trained networks. We instead analyze the spectral properties of randomized $\mathbf{W}$, look for large eigenvalues (λ_{trap}) that deviate from the MP baseline, and track them through extended training. Correlation Traps were first proposed in [13]; here we connect them to anti-grokking and to a practical output-space ablation diagnostic.

Self-averaging and overfitting. Our criterion connects to statistical-mechanics accounts of glassy learning, where poor generalization reflects sample-specific structure rather than a single stable rule [1, 4, 7, 8, 20]. The MP law gives a self-averaging baseline for randomized layer spectra, and Correlation Traps violate that baseline. Such traps can support a non-self-averaging generalization error through localization, where a small coordinate set retains $O(1)$ variance under subsampling, or through condensation, where a dominant spectral mode carries macroscopic variance. The corresponding physics analogies are Anderson localization, Bose–Einstein condensation, and Curie–Weiss mean-field magnetization [2, 5, 6, 22].

3. Method: MP Baseline and Correlation Traps

For a weight matrix $\mathbf{W} \in \mathbb{R}^{N\times M}$, define the layer covariance and empirical spectral density (ESD)

$$\mathbf{X} = \frac{1}{N}\mathbf{W}^\top\mathbf{W}, \qquad \rho_{\text{emp}}(\lambda) = \frac{1}{M}\sum_{i=1}^{M}\delta(\lambda-\lambda_i). \tag{1}$$

If entries of $\mathbf{W}$ are i.i.d. and well behaved, then in the limit $N, M \to \infty$ with $Q = N/M \geq 1$ fixed, ρ_{emp} converges to the MP density [11],

$$\rho_{\text{MP}}(\lambda) = \frac{Q}{2\pi\sigma^2}\frac{\sqrt{(\lambda_+-\lambda)(\lambda-\lambda_-)}}{\lambda}\mathbf{1}_{[\lambda_-,\lambda_+]}(\lambda), \qquad \lambda_\pm = \sigma^2(1\pm Q^{-1/2})^2. \tag{2}$$

At finite N, the right edge λ_+ lives within the scale of Tracy–Widom fluctuations. The MP bulk is consistent with **self-averaging**: in the large-N limit, the ESD concentrates onto a deterministic law, and the associated MP singular/eigenvectors are delocalized. Large spectral outliers represent **non-self-averaging** structure: a small number of directions dominate the statistics, and observables depending on them may fail to concentrate.

Recognizing that RMT requires matrix entries to be effectively independent and not too heavy-tailed, we randomize the layer elementwise, $\mathbf{W} \mapsto \mathbf{W}^{\text{rand}}$. Under the randomized null, the ESD of $\mathbf{X}^{\text{rand}} = N^{-1}(\mathbf{W}^{\text{rand}})^\top\mathbf{W}^{\text{rand}}$ should be well fit by an MP distribution if the layer is self-averaging. We call an eigenvalue of $\mathbf{X}^{\text{rand}}$ a **Correlation Trap** when

$$\lambda_{\text{trap}} > \lambda_+^{\text{MP}} + \Delta_{\text{TW}}, \tag{3}$$

where Δ_{TW} denotes the scale of finite-size Tracy–Widom fluctuations. By BBP theory, an eigenvalue that separates beyond the edge is a structural outlier rather than a typical bulk fluctuation [3]. The

`WeightWatcher` tool detects traps automatically by shuffling entries, running an SVD, fitting the ESD to the MP law, and identifying right-edge outliers. The expanded randomized-ESD figures and algorithm table appear in Appendix C.

A trap is not automatically harmful. For a covariance-like matrix $C = \sum_\alpha \lambda_\alpha v_\alpha v_\alpha^\top$ and a linearized observable with sensitivity vector b,

$$\mathrm{Var}(A_b) = b^\top C b = \sum_\alpha \lambda_\alpha \langle b, v_\alpha \rangle^2. \tag{4}$$

A trap is dangerous when one term in this sum remains macroscopic. This can happen through *geometric localization*, where v is supported on a small coordinate set, or through *spectral condensation*, where one large eigenvalue dominates even if its eigenvector is diffuse. Thus the MP violation is a weight-only candidate non-self-averaging mode, but its behavioral effect must be tested.

4. Empirical Results: Traps Track Anti-Grokking

We study three standard grokking benchmarks, denoted MLP, MA, and GPT2. The first is a depth-3 ReLU MLP trained on a balanced MNIST subset containing 100 examples from each class, for a total of 1,000 training points. The network has width 200 in each hidden layer and is trained with AdamW, MSE loss on one-hot targets, learning rate 5×10^{-4}, and up to 10^7 optimization steps. The main runs use `WD=0`, with a `WD=0.01` control. The second benchmark, MA, is a small one-layer transformer trained on $x + y \bmod P$. The third, GPT2, follows the GrokkedTransformer synthetic composition task [21]: atomic facts (h, r, t) are queried as $(h, r) \mapsto t$, latent composition rules produce inferred two-hop facts $(h, r_1, r_2) \mapsto t$, and "test" measures held-out in-distribution compositional generalization over unseen inferred facts. Full details are in Appendices E–E.4.

Figure 2 in Appendix A gives the corresponding test-loss view of the same dynamics.

MLP. Figure 1(a) shows the full long-horizon trajectory for the MNIST MLP. The run exhibits the familiar pre-grokking → grokking transition: training accuracy saturates rapidly, test accuracy improves much later, and the model reaches a high-generalization regime around 10^6 steps. Continued optimization reveals a third phase. After grokking, test accuracy drops substantially again while training accuracy remains essentially perfect; the test loss also reaches a minimum and then worsens (Appendix Fig. 2). Trap count separates this regime from the earlier phases. During pre-grokking, when the model has already fit the training set but has not yet generalized, detected traps are effectively zero. During grokking, trap count remains near zero. It is only when late-stage anti-grokking begins that trap count rises sharply.

Weight decay and perturbation controls. A useful control is the MLP run with nonzero weight decay. Weight decay does not eliminate the three phases entirely, but it reduces both the number of traps and the extent of the late-stage test degradation (Appendix E.2). This supports the non-self-averaging interpretation: it is not just that the weights are smaller, but that Correlation Traps are suppressed. A further perturbation study rules out a simpler explanation that trap counts merely track global weight norm. Even when checkpoint perturbations change norm-based quantities across phases, the shuffled-spectrum trap count remains near zero in pre-grokking and grokking, and becomes nonzero only in anti-grokking (Appendix G).

Modular addition and GPT2. The same signature recurs outside image classification. In MA, all measured layers remain trap-free in pre-grokking and grokking, then develop traps in anti-grokking; the mean anti-grok trap count across analyzed layers is 4.88 ± 1.55. In the GPT2-style composition task, layer-wise traps appear as the model loses held-out compositional generalization after initially grokking. Thus, the phenomenon is not tied to one architecture or task: dense image classifiers, algorithmic transformers, and GPT-style sequence models show the same trap-onset pattern.

5. Trap Classification by JSD Ablation

RMT detects candidate traps, but a trap can be benign. Let M_θ be the original model and $M_{\theta \setminus k}$ the model after replacing trap k with a suitable random vector. For probe inputs $\mathbf{x}_p$, define

$$J_k(T) = \mathbb{E}_{\mathbf{x}_p}\left[D_{\mathrm{JS}}\left(\mathrm{softmax}(z_\theta(\mathbf{x}_p)/T) \,\middle\|\, \mathrm{softmax}(z_{\theta \setminus k}(\mathbf{x}_p)/T)\right)\right]. \tag{5}$$

For MA and GPT2 we use random-token probes; for MLP we use Gaussian probes matched to input statistics. Large $J_k(T)$ means trap removal changes model outputs even without labels. Figure 3 in Appendix B shows that, within a trained model, the trap-removal score tracks the downstream test-error change induced by trap ablation. Thus JSD is best interpreted as a within-model measure of trap activity rather than a universal score.

6. Broader Implications for Frontier-Scale Models

We applied the same shuffled-spectrum diagnostic to two open-weight reasoning models, `gpt-oss-20b` and `gpt-oss-120b` [17, 18]. Figure 10 in Appendix I reports layer-wise trap counts. These profiles do not by themselves prove harmful overfitting, but they suggest a practical screening use case: trap counts can flag checkpoints deserving closer evaluation before deployment or fine-tuning.

7. Conclusion and Limitations

In this study, we show that when a model overfits its training data, it leaves distinct signatures in the layer weights; we call these signatures *Correlation Traps*. We study such 'traps' in grokking because, when overfitting a model that is grokking, there is a natural negative control: pre-grokking. Before grokking, the model may fit the training data and fail on the test data, but it has not yet learned the generalizing rule. After grokking, the model has learned that rule and then loses it under continued optimization. This late-stage failure, which we call anti-grokking, was unexpected and appears to be largely missing from the grokking literature. The key result is that Correlation Traps appear in anti-grokking, not pre-grokking: two regimes that look similar in accuracy but differ in the weights.

We study three classic grokking experiments, but for an extended period: MNIST MLP, Modular Addition (MA), and GPT2. In all three, anti-grokking produces outliers (λ_{trap}) that we can detect using Random Matrix Theory (RMT). The diagnostic is simple and weight-only: shuffle each layer entry-wise, fit a Marchenko–Pastur (MP) bulk to the randomized covariance spectrum, and count right-edge outliers beyond the MP/TW edge. We then use JSD ablation to test whether these detected traps affect model behavior. Correlation Traps can detect the onset and the extent of the kind of classic overfitting we observe in anti-grokking. Such traps provide a unique diagnostic to detect overfitting in models where test and training accuracy may be unavailable and only weights are present, such as, most notably, foundation-scale LLMs.

References

[1] Daniel J. Amit, Hanoch Gutfreund, and Haim Sompolinsky. Spin-glass models of neural networks. *Physical Review A*, 32(2):1007–1018, 1985. doi: 10.1103/PhysRevA.32.1007.

[2] Philip W. Anderson. Absence of diffusion in certain random lattices. *Physical Review*, 109(5): 1492–1505, 1958. doi: 10.1103/PhysRev.109.1492.

[3] Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643–1697, 2005.

[4] Siegfried Bös. Statistical mechanics approach to early stopping and weight decay. *Physical Review E*, 58(1):833–847, 1998. doi: 10.1103/PhysRevE.58.833.

[5] Satyendra Nath Bose. Plancks Gesetz und Lichtquantenhypothese. *Zeitschrift für Physik*, 26 (1):178–181, 1924. doi: 10.1007/BF01327326.

[6] Albert Einstein. Quantentheorie des einatomigen idealen Gases. *Sitzungsberichte der Preussischen Akademie der Wissenschaften*, pages 3–14, 1925.

[7] E. Gardner. The space of interactions in neural network models. *Journal of Physics A: Mathematical and General*, 21(1):257–270, 1988. doi: 10.1088/0305-4470/21/1/030.

[8] John J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982. doi: 10.1073/pnas.79.8.2554.

[9] Jaerin Lee, Bong Gyun Kang, Kihoon Kim, and Kyoung Mu Lee. Grokfast: Accelerated grokking by amplifying slow gradients, 2024. URL https://arxiv.org/abs/2405.20233.

[10] Ziming Liu, Ouail Kitouni, Niklas S. Nolte, Eric J. Michaud, Max Tegmark, and Mike Williams. Towards understanding grokking: An effective theory of representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 34651–34663. Curran Associates, Inc., 2022. URL https://arxiv.org/abs/2205.10343.

[11] Vladimir A. Marchenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, 72(114)(4):507–536, 1967.

[12] Charles H. Martin. WeightWatcher: Analyze Deep Learning Models without Training or Data. https://github.com/CalculatedContent/WeightWatcher, 2018-2024. Version 0.7.5.5 used in this study. Accessed May 12, 2025.

[13] Charles H. Martin and Christopher Hinrichs. SETOL: A semi-empirical theory of (deep) learning. *arXiv preprint arXiv:2507.17912*, 2025. URL https://arxiv.org/abs/2507.17912.

[14] Charles H. Martin and Michael W. Mahoney. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *Journal of*

Machine Learning Research, 22(1):165, January 2021. URL http://jmlr.org/papers/v22/20-410.html.

[15] Charles H. Martin, Tian Peng, and Michael W. Mahoney. Predicting trends in the quality of state-of-the-art neural networks without access to training or testing data. *Nature Communications*, 12: 4122, jul 2021. doi: 10.1038/s41467-021-24025-8. URL https://doi.org/10.1038/s41467-021-24025-8.

[16] Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability, 2023. URL https://arxiv.org/abs/2301.05217.

[17] OpenAI. Introducing gpt-oss. https://openai.com/index/introducing-gpt-oss/, August 2025. Accessed: 2026-03-28.

[18] OpenAI. gpt-oss-120b & gpt-oss-20b model card. https://openai.com/index/gpt-oss-model-card/, August 2025. Accessed: 2026-03-28.

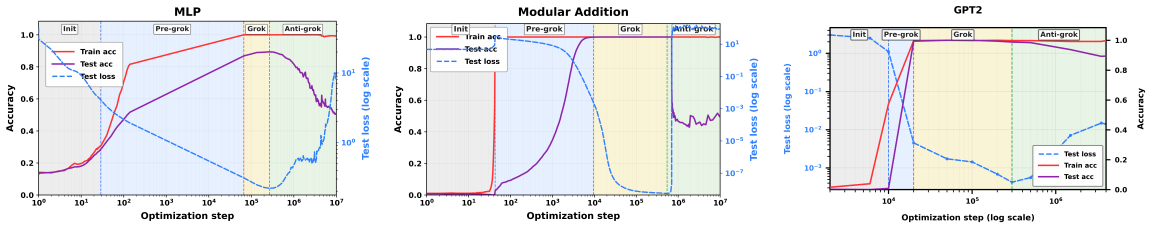
[19] Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets, 2022. URL https://arxiv.org/abs/2201.02177.

[20] H. Sebastian Seung, Haim Sompolinsky, and Naftali Tishby. Statistical mechanics of learning from examples. *Physical Review A*, 45(8):6056–6091, 1992. doi: 10.1103/PhysRevA.45.6056.

[21] Boshi Wang, Xiang Yue, Yu Su, and Huan Sun. Grokked transformers are implicit reasoners: A mechanistic journey to the edge of generalization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://arxiv.org/pdf/2405.15071.

[22] Pierre Weiss. L'hypothèse du champ moléculaire et la propriété ferromagnétique. *Journal de Physique Théorique et Appliquée*, 6(1):661–690, 1907. doi: 10.1051/jphystap:019070060066100.

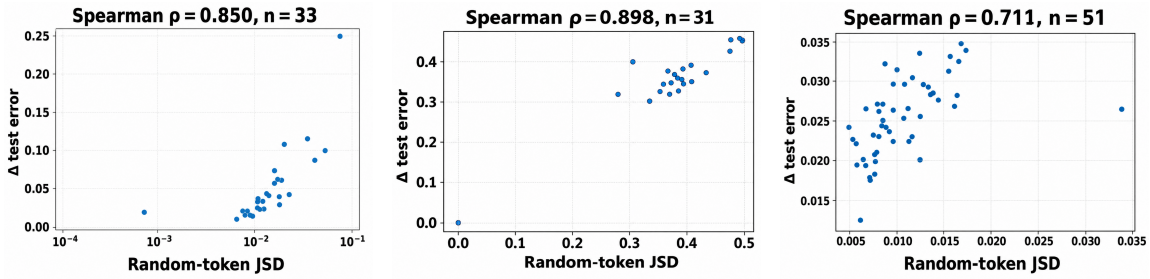
Appendix A. Training-Loss Curves



(a) MLP (b) MA (c) GPT2

Figure 2: **Grokking dynamics and test loss.** Train accuracy, test accuracy, and test loss are shown for the MLP, MA, and GPT2 experiments. Test loss reaches a good-generalization regime and later worsens under continued optimization.

Appendix B. JSD Ablation Diagnostic Plots



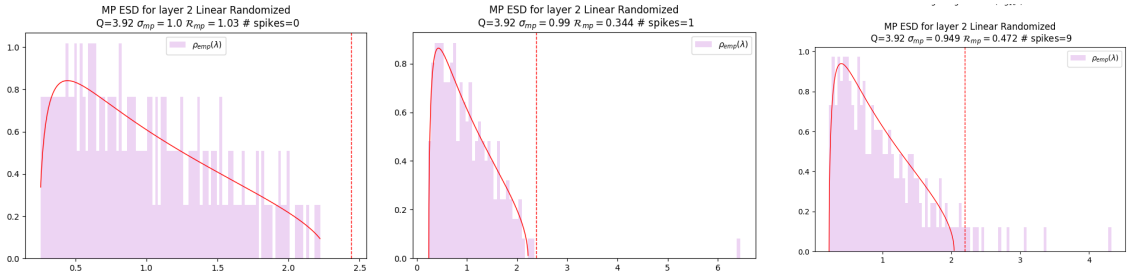
(a) MLP (b) MA (c) GPT2

Figure 3: **JSD diagnostic ablation test.** Trap-removal score $J_k(T = 1)$ versus change in test error for selected MLP, MA, and GPT2 anti-grokking checkpoints. Within a trained model, larger JSD indicates a more behaviorally active trap.

Appendix C. Expanded Random-Matrix Diagnostic

The open-source `WeightWatcher` tool implements random-matrix-based analyses of neural-network layers. Here it is used to examine individual layer spectral densities. The key object is the covariance $\mathbf{X} = N^{-1}\mathbf{W}^\top\mathbf{W}$ and its ESD. If entries are independent and well behaved, the ESD follows the MP law in the large-aspect-ratio limit, and the associated singular/eigenvectors are delocalized with components of order $1/M$ up to fluctuations.

The diagnostic is nonparametric and practical: it does not determine the exact mechanism producing the outlier, but it identifies a candidate failure of self-averaging. Some traps are harmful and some benign, so detection is followed by the JSD ablation test.



(a) (b) (c)

Figure 4: **Randomized ESDs, MP fits, and Correlation Traps.** (a) A self-averaging randomized layer follows the MP bulk with no large outliers. (b,c) Trapped randomized spectra from Layer 2 of the MLP show separated spikes beyond the MP bulk.

Table 1: Trap-detection algorithm used to compute trap-count curves.

Step	Procedure
1	Take a trained layer $\mathbf{W} \in \mathbb{R}^{N\times M}$ at a saved checkpoint.
2	Shuffle entries W_{ij} elementwise to form $\mathbf{W}^{\mathrm{rand}}$.
3	Compute SVD/eigenvalues of $\mathbf{X}^{\mathrm{rand}} = N^{-1}(\mathbf{W}^{\mathrm{rand}})^\top \mathbf{W}^{\mathrm{rand}}$.
4	Fit the ESD $\rho_{\mathrm{emp}}(\lambda)$ to the MP law.
5	Count eigenvalues $\lambda_{\mathrm{trap}} > \lambda_+^{\mathrm{MP}} + \Delta_{\mathrm{TW}}$.

Appendix D. Traps as Signs of Non-Self-Averaging

Self-averaging is the statistical-mechanics analogue of concentration. An observable A_N self-averages when its relative fluctuations vanish,

$$\frac{\mathrm{Var}(A_N)}{\mathbb{E}[A_N]^2} \to 0. \tag{6}$$

For empirical risk $R_n = n^{-1}\sum_i L_i$,

$$\mathrm{Var}(R_n) = \frac{1}{n^2}\mathbf{1}^\top \Sigma_n \mathbf{1}, \qquad (\Sigma_n)_{ij} = \mathrm{Cov}(L_i, L_j). \tag{7}$$

Concentration requires no macroscopic sample-specific covariance mode. More generally, for $C = \sum_\alpha \lambda_\alpha v_\alpha v_\alpha^\top$ and sensitivity b, $\mathrm{Var}(A_b) = \sum_\alpha \lambda_\alpha \langle b, v_\alpha\rangle^2$. A trap is dangerous when one term remains macroscopic.

Localization. A trap mode may be localized: $\sum_{i\in S} v_i^2 \approx 1$ with $|S| \ll M$. If an observable has non-negligible projection on this support, it averages over only a small effective number of coordinates, so fluctuations need not decay at the usual rate. In practice, traps often contain a significant fraction of the top 5% mass of W_{ij}, but localization alone provides only limited predictive signal for harmfulness.

Condensation. A trap may also be diffuse but spectrally dominant. If $\lambda_{\max}$ is much larger than the rest, then $\lambda_{\max}\langle b, v_1\rangle^2$ can dominate even when v_1 is spread out. Thus self-averaging can fail through geometric localization, spectral condensation, or both.

D.1. Self-Averaging, Concentration, and the MP Null

In statistical mechanics, one studies sample-to-sample fluctuations of macroscopic quantities such as free-energy density, magnetization, or overlap. In learning theory, the same idea appears as concentration of measure: an observable A_N self-averages if $A_N - \mathbb{E}A_N \to 0$ in probability, and a sufficient L^2 criterion is $\mathrm{Var}(A_N) \to 0$.

The spin-glass interpretation of overfitting predates modern concentration language. Hopfield and Amit–Gutfreund–Sompolinsky models showed that associative memories develop glassy phases with sample-specific metastable states once interference is strong. Gardner's program studied the geometry of interactions compatible with many patterns, and statistical-mechanics analyses of learning interpreted poor generalization as movement into glassy, sample-dependent states, with weight decay or early stopping keeping the system away from that regime.

For covariance-type random matrices with independent, well-behaved entries, the ESD concentrates around MP. The vector statement is equally important: eigenvectors are delocalized and random-looking. A shuffled layer with a separated edge eigenvalue or localized top vector has departed from this null. The relevant benchmark is the fitted MP edge plus TW-scale fluctuations; BBP theory motivates treating cleanly separated eigenvalues as structural outliers.

The diagnostic uses an entry-wise shuffle of the learned multiset of entries. The scientific point does not depend on with- versus without-replacement sampling conventions. What matters is localization of the trap vectors and spectral size of the associated outliers, not a specific asymptotic tail model.

Appendix E. Experimental Setup and Additional Notes

E.1. MLP Experimental Setup

We train an MLP on a balanced MNIST subset of 1,000 training points, with 100 samples from each class. The main 10^7-step experiment used a single NVIDIA Quadro P2000 GPU and took approximately 11 hours; a considerable fraction of wall-clock time came from checkpointing and saving measurements.

E.2. MLP Control Experiment with Weight Decay

The primary results use zero weight decay to isolate long-horizon optimization dynamics from explicit norm regularization. A control with $\texttt{WD} = 0.01$ uses the same architecture, data, and optimizer. Weight decay suppresses trap growth and reduces late-stage collapse, but does not eliminate the phenomenon entirely.

E.3. Modular Addition Experiment

The modular-addition benchmark uses a small one-layer transformer trained far beyond first generalization.

Trap counts are effectively zero in pre-grokking and grokking and rise sharply across many layers in anti-grokking.

Table 2: MLP experimental hyperparameters.

Parameter	Value
Network	Fully connected MLP
Depth	3 linear layers (Input $\to$ Hidden1 $\to$ Hidden2 $\to$ Output)
Width	200 hidden units per hidden layer
Activation	ReLU
Input/output size	784 / 10
Initialization	Default PyTorch weights and biases, scaled by 8.0
Dataset	MNIST
Train/test points	1,000 / 10,000
Batch size	200
Loss	MSE with one-hot targets
Optimizer	AdamW
Learning rate	5×10^{-4}
Weight decay	0.0 main; 0.01 control
AdamW $\beta_1, \beta_2, \epsilon$	$0.9, 0.999, 10^{-8}$
Steps	10^7
Data type	`torch.float64`
Seed	0
Framework/tooling	PyTorch; `WeightWatcher` v0.7.5.5

Table 3: Average number of Correlation Traps in MLP with and without weight decay.

Setting, Layer	Pre-Grokking	Grokking	Anti-Grokking
`WD=0`, FC1	0 ± 0	0 ± 0	7.5 ± 5.6
`WD=0`, FC2	0 ± 0	0 ± 0	1 ± 0
`WD>0`, FC1	0	0	2.0 ± 0.0
`WD>0`, FC2	0	0	1.0 ± 0.0

E.4. GPT2 Experimental Setup

We evaluate the shuffled-spectrum diagnostic on the synthetic composition benchmark from Grokked-Transformer. The task is generated from a random knowledge graph of atomic one-hop facts and latent two-hop composition rules.

For spectral and ablation analyses, `WeightWatcher` is applied to transformer block matrices while excluding the positional embedding layer `wpe`, since `wpe` is an embedding table rather than a standard learned linear map.

Appendix F. Mechanistic Case Study: MLP Anti-Grokking

The randomized trap itself is a diagnostic of atypical amplitude disorder. To interpret the failure mode, we map the trap mode back to original unshuffled layer coordinates. The trap indicates that a direction has moved into an overfit sector of weight space; the prototype-like structure appears when that direction is expressed in the coordinates of the original layer.

Table 4: Modular addition architecture.

Hyperparameter	Value
Layers	1
Model dimension d_{model}	128
MLP hidden size d_{mlp}	512
Heads $\times$ head dim	4×32
Context length n_{ctx}	3
Activation	ReLU
LayerNorm	disabled, `use_ln=False`
Vocabulary size	114 (= `equals_token` + 1)

Table 5: Modular addition phase summary.

Phase	Train Acc.	Test Acc.
Pre-grok	1.0 ± 0.0	0.40 ± 0.28
Grok	1.0 ± 0.0	0.97 ± 0.02
Anti-grok	1.0 ± 0.0	0.68 ± 0.11

F.1. Norm-Based Prototype Collapse

Figure 5 shows that the anti-grokking MLP no longer uses stable digit geometry in the usual sense. On original MNIST, the confusion matrix remains structured, but after each image is independently pixel-shuffled, predictions collapse toward a small number of classes, especially class 5. Matched Gaussian probes, which contain no class information or coherent strokes, again produce almost exclusively class 5 with nearly constant confidence. Thus late-stage predictions are controlled by $\|\mathbf{x}\|_2$-like input magnitude rather than semantic digit shape.

F.2. Intervening on the Trap Direction

Figure 7 shows that the structural interpretation is not merely visual. Replacing an FC1 trap direction with a matched Gaussian random vector changes the confusion structure and weakens the prototype bias toward class 5, showing that the overfit sector identified through the trap is behaviorally active.

When a localized trap is removed, it typically reappears under continued training. The same data, objective, and optimizer continue to favor that condensed direction, consistent with the trap being an emergent finite-size instability rather than a static artifact of one checkpoint.

F.3. Leading Eigenvectors, Receptive Fields, and Structural Outliers

For $\mathbf{W} \in \mathbb{R}^{N \times M}$, define $\mathbf{X}(\mathbf{W}) = N^{-1}\mathbf{W}^\top\mathbf{W}$. If $\mathbf{W} = U\mathbf{\Sigma}V^\top$, then eigenpairs of $\mathbf{X}(\mathbf{W})$ are $(\sigma_k^2/N, \mathbf{v}_k)$, so the leading right singular vector identifies the dominant input-space direction. In the MLP, the first-layer leading vector is directly visualizable in pixel coordinates. Rows of $\mathbf{W}_1$ selected by leading-vector localization evolve from noise-like patterns to digit-like receptive fields in anti-grokking, consistent with sharply localized prototype directions.

Table 6: Modular addition trap counts by layer and phase.

Layer	Pre-grok	Grok	Anti-grok
embed.embed	0.00	0.00	7
blocks.0.attn.W_Q	0.00	0.00	3
blocks.0.attn.W_K	0.00	0.00	5
blocks.0.attn.W_V	0.00	0.00	3
blocks.0.attn.out_proj	0.00	0.00	4
blocks.0.mlp.fc1	0.00	0.00	5
blocks.0.mlp.fc2	0.00	0.00	7
unembed.unembed	0.00	0.00	5
Mean±Std	$\mathbf{0.00 \pm 0.00}$	$\mathbf{0.00 \pm 0.00}$	$\mathbf{4.88 \pm 1.55}$

Table 7: GPT2 synthetic composition dataset.

Dataset parameter	Value
Dataset variant	`composition.2000.200.12.6`
Entities / relations	2000 / 200
Atomic outgoing facts per entity	20
Atomic fact/query form	(h, r, t), queried as $(h, r) \mapsto t$
Composition rule	$(h, r_1, m), (m, r_2, t) \Rightarrow (h, r_1, r_2) \mapsto t$
Atomic partition	$\text{atomic}_{\text{ID}} \cup \text{atomic}_{\text{OOD}}$
OOD atomic fraction	approximately 5%
Training inferred downsampling	$12.6 \times \lvert\text{atomic}_{\text{ID}}\rvert$

Table 8: GPT2 train/test split definitions.

Split	In training?	Role
id_atomic	Yes	ID atomic memorization/control
ood_atomic	Yes	OOD atomic memorization/control
$\text{train\_inferred}_{\text{ID}}$	Yes	Main "train" curve
$\text{test\_inferred}_{\text{ID}}$	No	Main "test" curve
$\text{test\_inferred}_{\text{OOD}}$	No	Stronger systematicity diagnostic

Table 9: GPT2 model, training, and evaluation configuration.

Hyperparameter	Value
Architecture	GPT2-style decoder-only transformer
Initialization	From scratch
Layers	8
Vocabulary	Synthetic entity, relation, and answer-delimiter tokens
Maximum sequence length	10
Batch size	512
Optimizer	AdamW
Learning rate / weight decay	10^{-4} / 0.001
Schedule / precision	Constant with warmup / mixed precision
Training steps	1.5×10^6
Evaluation protocol	Exact-match generation
Correctness	Generated answer token matches target before closing delimiter

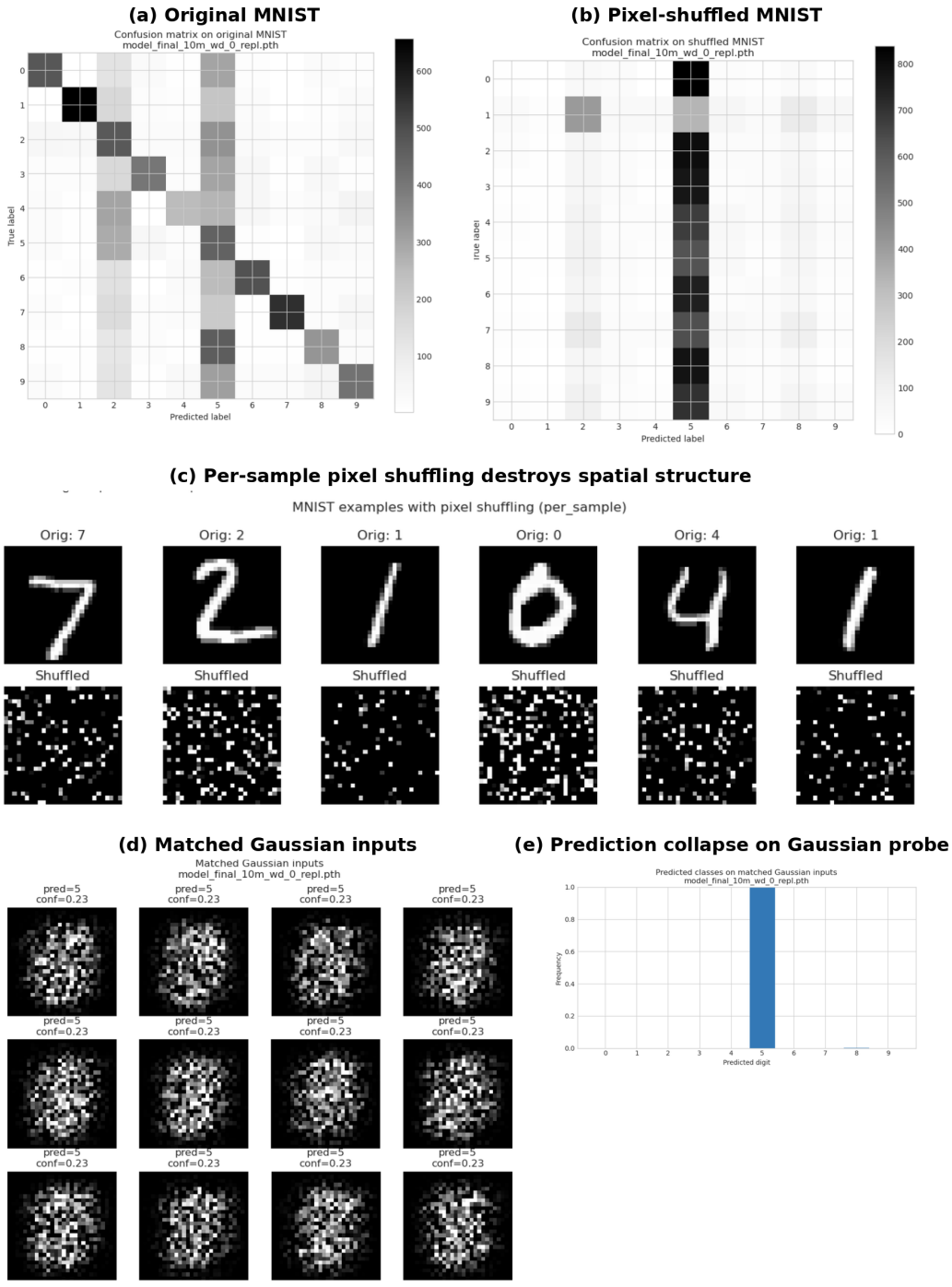
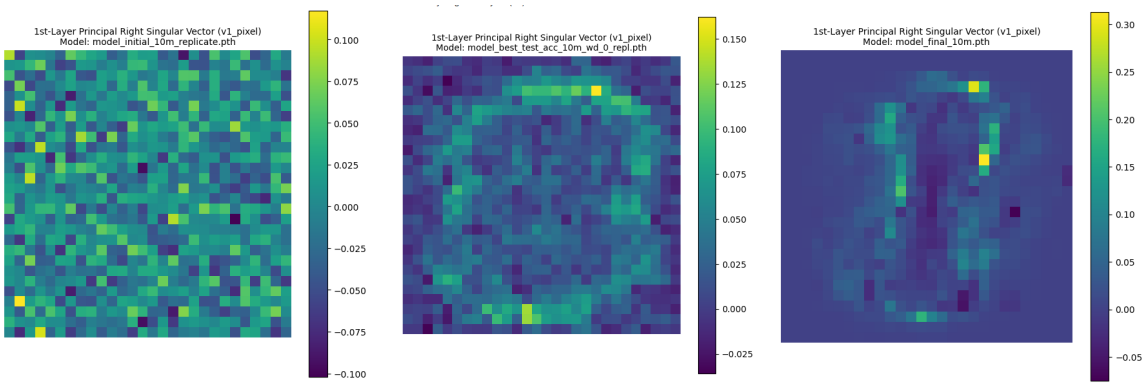


Figure 5: Evidence for norm-based prototype collapse in the anti-grokking MLP. Pixel-shuffled and matched Gaussian probes destroy digit structure but preserve low-order global statistics; predictions collapse toward a small set of labels.



(a) Pre-grokking (b) Grokking (c) Anti-grokking

Figure 6: Prototype-like directions associated with trapped layers. The largest right singular vector of $\mathbf{W}_1$ evolves from noise, to a smooth global template, to a localized prototype-like image.

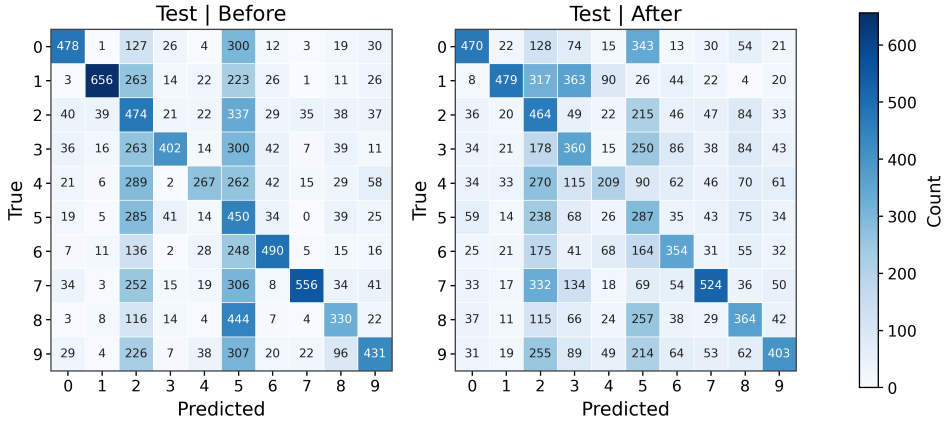


Figure 7: Trap intervention in anti-grokking. The full model shows a strong prototype bias; after replacing the FC1 trap direction, the bias is substantially reduced.

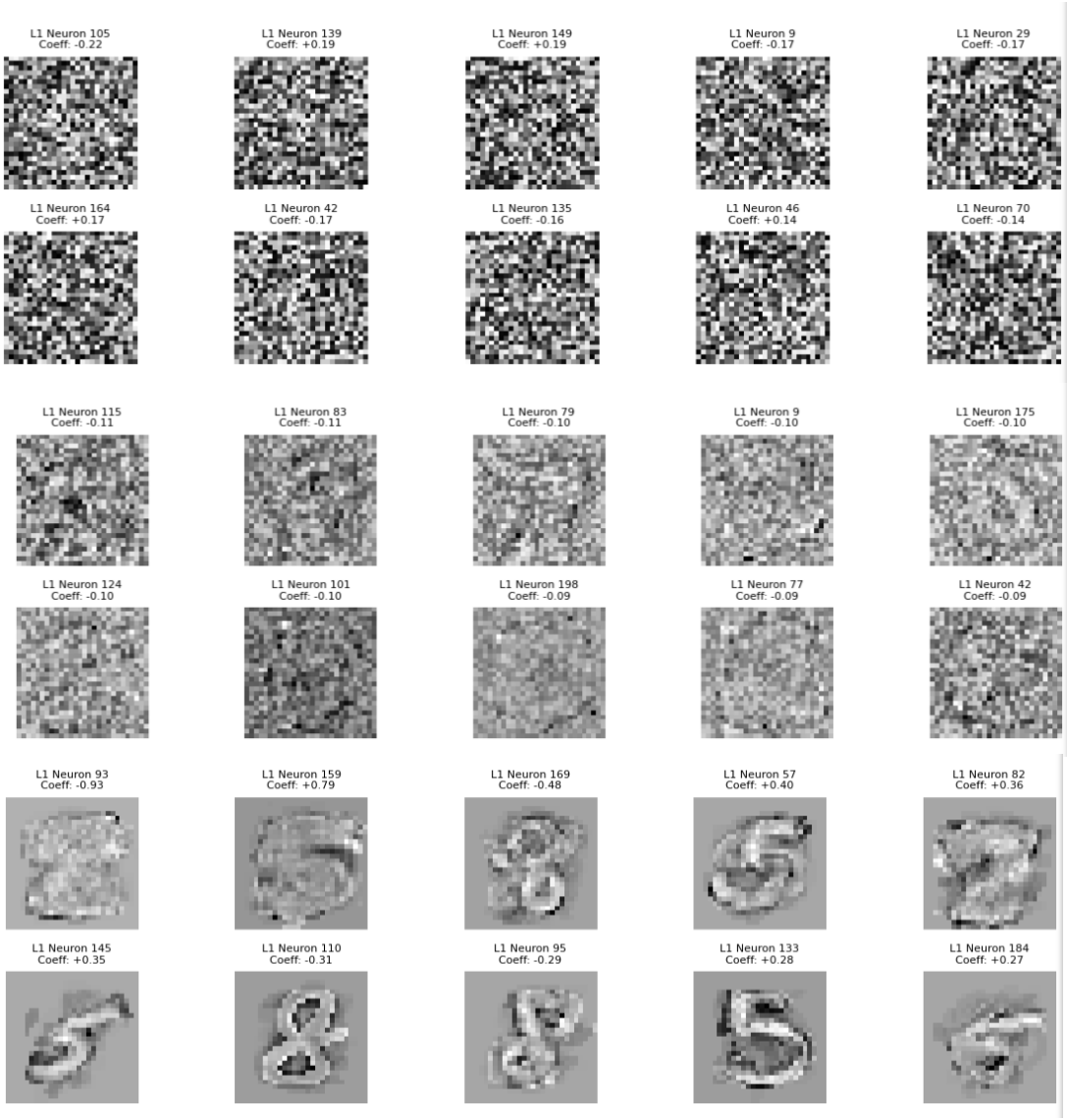
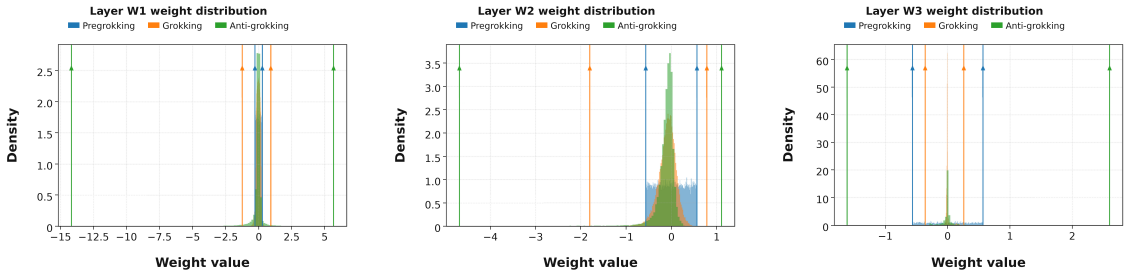


Figure 8: Rows of $\mathbf{W}_1$ selected by leading-vector localization. The selected receptive fields are noise-like in pre-grokking, remain diffuse around peak generalization, and become recognizable digit templates in anti-grokking.

F.4. Why Not Just Fit Heavy Tails?

A threshold on the largest entries is ad hoc, scale-dependent, and cannot distinguish diffuse large-weight growth from a genuinely condensed harmful mode. Direct tail fitting is also unstable: histograms can look Laplacian, power-law-like, or mixed depending on the regime and cutoff. The MP law gives a nonparametric self-averaging baseline for the randomized layer. The mechanism does not require exponent < 2; empirically, traps can be driven by several large or moderately large entries whose combined squared mass produces a localized direction.



(a) $\mathbf{W}_1$ (b) $\mathbf{W}_2$ (c) $\mathbf{W}_3$

Figure 9: Weight distributions with extreme coordinates across checkpoints. Anti-grokking introduces structural outliers into the weights of all three layers.

Appendix G. MLP Perturbation Study

This check rules out the possibility that trap counts merely track weight scale. We perturb MLP checkpoints so global norm-based quantities move across phases. Even under this perturbation, trap counts remain near zero in pre-grokking and grokking and become nonzero only in anti-grokking. Phase boundaries are pre-grok 10^2–5×10^4, grok 5×10^4–5×10^5, and anti-grok $> 5 \times 10^5$ steps.

Table 10: Accuracy by phase under perturbation.

Phase	Train Accuracy	Test Accuracy
Pre-grok	0.7298 ± 0.0408	0.4926 ± 0.0187
Grok	1.0000 ± 0.0000	0.8879 ± 0.0051
Anti-grok	1.0000 ± 0.0000	0.6184 ± 0.1086

Table 11: Global ℓ_2 norm by phase under perturbation.

Phase	WeightNorm
Pre-grok	17.52 ± 0.00
Grok	16.80 ± 0.22
Anti-grok	27.64 ± 8.39

Table 12: Layer-wise trap statistic by phase under perturbation.

Layer	Pre-grok	Grok	Anti-grok
1	0.00	0.00	7.00
3	0.00	0.00	1.17
5	0.00	0.00	5.50
Mean±Std	$\mathbf{0.00 \pm 0.00}$	$\mathbf{0.00 \pm 0.00}$	$\mathbf{4.56 \pm 3.03}$

Appendix H. Robustness to Random Seeds and Shuffle Randomization

These checks support two paper-level claims: the three-phase trajectory is not a single-run artifact, and the shuffled-spectrum diagnostic is not driven by one lucky permutation.

Table 13: Qualitative robustness across random seeds.

Phase	Observed behavior	Interpretation
Pre-grokking	Train accuracy rises while test accuracy remains low; trap counts remain approximately zero	Stable trap-free regime
Grokking	Test accuracy improves sharply; shuffled spectra remain MP-like with few/no traps	Stable generalization regime
Anti-grokking	Test accuracy declines while trap counts become clearly nonzero	Stable trapped overfitting regime

Table 14: Qualitative robustness to repeated entry-wise randomization.

Phase	Effect of repeated shuffles	Interpretation
Pre-grokking	Microscopic variation around zero traps	Remains trap-free at paper level
Grokking	Same qualitative behavior; no stable right-edge outliers	Remains trap-free/nearly trap-free
Anti-grokking	Count fluctuates slightly but stays clearly nonzero	Anomaly is structural rather than one-shuffle artifact

Appendix I. Frontier-Scale Screening

We applied the shuffled-spectrum diagnostic to OpenAI `gpt-oss-20b` and `gpt-oss-120b`. Figure 10 reports layer-wise trap counts. These profiles do not by themselves prove harmful overfitting, but they motivate a practical screening use case: trap counts can identify layers/checkpoints for closer task-specific and probe-based evaluation.

Appendix J. Expanded Conclusion and Limitations

The main lesson is that grokking is not always the end of training. A model can first learn a rule that generalizes, and then lose that generalization under continued optimization. We call this late-stage

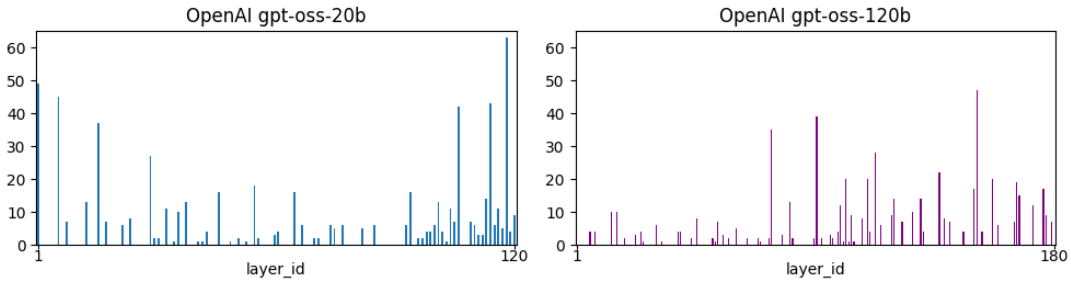


Figure 10: Layer-wise Correlation Trap profiles in GPT-OSS 20B and 120B.

failure mode *anti-grokking*: specifically, post-generalization, test accuracy degrades, while training accuracy remains high. Across MNIST MLP, Modular Addition (MA), and GPT2-style grokking experiments, this transition is marked by the appearance of outliers (λ_{trap}) in the eigenspectrum of one or more randomized, layer-wise weight matrices. We call these outliers *Correlation Traps*.

The term Correlation Trap is operational. We call these outliers Correlation Traps because they identify latent structures that appear to "trap" the correlations, causing the model to overfit its data, reducing the test accuracy of the model, but not necessarily the training accuracy. And as the overfitting gets worse, more traps appear. Moreover, removing them can even hurt the model's test and training accuracy, apparently removing the correlations the model had learned.

Correlation Traps reveal a structural difference between two superficially similar regimes. In both pre-grokking and anti-grokking, training accuracy can be high while test accuracy is poor. But the weight spectra are different. During pre-grokking, trap counts remain near zero. During successful grokking, they also remain absent or nearly absent. They appear only during anti-grokking; thus, traps do not merely detect memorization, low test accuracy, or high training accuracy. They detect overfitting that arises during anti-grokking.

The diagnostic is simple and weight-only. It does not require training data, test data, gradients, optimizer state, or access to the training pipeline. It only requires a checkpoint. For each layer, we shuffle the weight matrix entry-wise, form the randomized covariance spectrum, fit a Marchenko–Pastur bulk, and count right-edge outliers beyond the fitted MP/TW edge. The MP/TW law provides a self-averaging baseline. A Correlation Trap is a separated spectral mode that violates this baseline, and may be harmful or benign.

We do not claim that every form of overfitting must produce Correlation Traps. The claim here is sharper: in long-horizon grokking, trap onset tracks anti-grokking while pre-grokking provides the trap-free negative control. Using the JSD diagnostic trap ablation test, traps can be identified as harmful or benign for a given model. These results suggest some forms of harmful overfitting leave detectable signatures in the layer weight matrices of seemingly well-trained models. The next step is to use this diagnostic during training and fine-tuning, and to test whether harmful Correlation Traps can be suppressed, removed, or regularized without damaging useful learned structure.

Appendix K. Reproducibility Artifacts

The supplemental material accompanying this submission should include an anonymized artifact bundle for reproducing the main experiments: (i) training scripts for the MLP and modular-addition

benchmarks; (ii) analysis scripts for shuffled-spectrum trap counts, KS tests, and figure generation; (iii) checkpoints or scripts to regenerate representative checkpoints; and (iv) a README with software environment, package versions, and commands. The datasets and tools are public: MNIST is used for the MLP benchmark, and trap analysis uses `WeightWatcher`.

Appendix L. Assets, Licenses, and Attribution

The paper uses existing public assets. The MLP benchmark uses MNIST. The shuffled-spectrum analysis uses `WeightWatcher` version v0.7.5.5 under Apache License 2.0. The exploratory frontier-model analysis uses public `gpt-oss-20b` and `gpt-oss-120b` open-weight releases, made available under Apache License 2.0 and the corresponding OpenAI usage policy. We do not introduce a new dataset.